Webディレクトリ拡張の自動化手法

風間一洋† 原田昌紀† 佐藤進也†

Web の急速な成長にともない，Web ディレクトリをつねに最新の状態に保つことは困難になりつつある．本論文では，Web ロボットで収集した大量の Web ページ集合を使用して，Web ディレクトリを自動的に拡張する手法を提案する．本手法は，テキストの類似性ではなく，ハイパーリンク構造を基にする．まず最初に，共参照解析によって，各カテゴリに関連した権威ある Web サイトを発見する．次に，発見した Web サイトに対する説明文を，それにリンクしている Web ページから抽出する．検証用システムとして開発した ODIN ディレクトリでは，Open Directory が提供している Japanese カテゴリの下のすべてのカテゴリに対して拡張手法を適用し，本手法が 700 以上の詳細なカテゴリに対して正確に妥当かつ権威ある Web サイトを検出できることを示す．さらに，ODIN ディレクトリを一般公開し，不特定多数の利用者の行動を解析することで，拡張したデータが元データと同等にアクセスされていることを示す．

Automated Method for Web Directory Expansion

Kazuhiro Kazama,† Masanori Harada† and Shin-ya Sato†

With the rapid growth of the Web, it is a challenging issue to maintain web directories up-to-date. In this paper, we propose a method to expand a web directory automatically by using huge amount of web pages collected by a web robot. It is not a content-based approach, but is a hyperlink-based approach. It consists of two steps. First, we find authoritative web sites relevant to each category by co-citation analysis. Second, we extract descriptions of found web sites from web pages linking to them. We developed a testbed system named “the ODIN Directory” and expanded all of the categories under the Japanese category of the Open Directory automatically. Our experiments showed that our method could find accurately relevant and authoritative web sites for each category while the Japanese category consists of more than 700 detailed categories. We also put the ODIN Directory on the public web site and confirmed that anonymous users accessed the expanded part of the web directory equally to the original part by user behavior analysis.

1. はじめに

Web 情報検索システムは，Web サーチエンジンと Web ディレクトリの 2 種類に大きく分類できる．前者の例は Google であり，Web ロボットで収集した膨大な量の Web ページの索引を作成し，検索に使用する．後者の例は Yahoo!であり，トピックごとに階層的に分類された Web サイトを閲覧または検索に使用する．Web ディレクトリは Web サーチエンジンに対して次のような利点を持つ．

(1) 利用者は，検索質問を入力せずに情報を探すことができる．
(2) 編集者が審査し，適切と判断した Web サイトだけが登録されている．
(3) 編集者が記述した，短く的確な説明文が Web サイトにつけられている．
(4) Web ページ単位ではなく，Web サイト単位で探すことができる．

なお，Web 空間には，利用者の要求に適合した Web サイトは大量に存在するが，その質は千差万別であり，内容が不完全だったり，誤りを含んでいることさえもある．利用者にとって望ましいのは，その中から，内容が正確で信頼できる少数の Web サイトだけを閲覧できることである．このような Web サイトを，特に**権威ある Web サイト**と呼ぶ[1]．つまり，Web ディレクトリの有用性を確保するためには，編集者が権威ある Web サイトだけを選択し，それを適切なカテゴリに分類し，簡潔で適切な説明文をつけることが重要である．

しかし，Web サイトは頻繁に誕生または移動するために，Web ディレクトリをつねに最新の状態にしておくことは本質的に困難であり，実際に致命的な遅

† NTT 未来ねっと研究所
NTT Network Innovation Laboratories

延が生じている．さらに，登録・更新作業は人手によるので，多くの熟練した編集者と多大な費用が必要になる．

本論文では，編集者が見逃した有用な関連 Web サイトを自動的に発見して，その説明文と一緒に追加することにより，Web サイトの登録・更新作業の一部を自動化する手法を提案する．このように，Web ディレクトリを，登録数，登録されている Web サイトの有用性，および有用な Web サイトの登録遅延と更新遅延の短縮などの点で改善する手法を，**Web ディレクトリ拡張**と呼ぶ．

2 章では，Web ディレクトリ拡張の概要について述べる．本手法は，各カテゴリに対する関連 Web サイト発見と，関連 Web サイトの説明文発見の 2 段階に大きく分類できる．3 章では，共参照解析を用いて権威ある関連 Web サイトを発見する手法について述べる．ここで提案する Multi Co-citation アルゴリズムは，複数の Web サイトをアルゴリズムの起点とし，より多くの Web サイトと高い共参照関係にある Web サイトを高く評価することで，全体としてより良い適合度が得られるようにしたアルゴリズムである．4 章では，関連 Web サイトを参照している Web サイトの文の中から，パターンマッチングに基づいて最適な説明文を発見する手法について述べる．5 章では，実際に Open Directory のデータに本手法を適用して実験を行い，Open Directory のカテゴリに登録されている Web サイトを本手法で自動分類しなおした場合の分類精度，関連 Web サイトと抽出された説明文の適合性について評価する．6 章では，本手法を用いた ODIN ディレクトリを一般公開して得られた利用者の利用ログを分析した結果について述べる．7 章では，関連研究について述べる．8 章では，結論を述べる．

2. Web ディレクトリ拡張

2.1 Web ディレクトリ拡張の方針

本論文の Web ディレクトリ拡張手法は，次の 3 つの仮定に基づく．

(1) Web サイトの作成者が他の Web サイトをリンクする場合は，そこに閲覧者が訪問する何らかの価値があると考えている．

(2) （特にリンク集のような）Web ページでは，同一トピックの Web サイトを複数リンクする傾向がある．

(3) （特にリンク集のような）Web ページでは，リンクした Web サイトに Web サイトの作成者が書いた適切・簡潔な紹介文を付けることが多い．

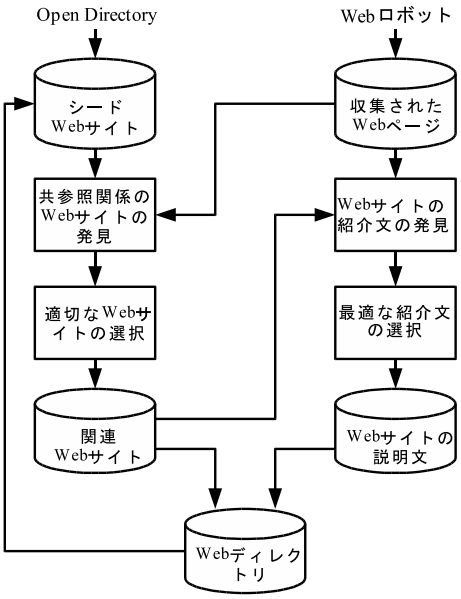

図 1 Web ディレクトリ拡張の概要
Fig. 1 The overview of Web directory expansion.

(1) から，被リンク数が多い Web サイトは Web ディレクトリに登録する価値があると推測し，被リンク数が多いものを優先する．(2) から，カテゴリに登録されている Web サイトと**共参照（co-citation）**関係にある Web サイトは何らかの関連性を持つと推測し，より高い共参照度を持つものを探す．(3) から，ある Web サイトをリンクした Web ページ集合に，その紹介文が存在すると推測し，Web ディレクトリの説明文に最も適したものを探す．

ただし，本手法では Web ページの内容には深くは踏み込まない．Web ページの内容に基づいたテキスト分類技術およびテキスト要約技術については多くの研究が存在するが，一般に Web ページはさまざまな形式（例，書き言葉，話し言葉，俗語など），さまざまな言語（例，英語，日本語，中国語など）で記述され，新しい用語や言い回しも頻繁に誕生しているので，プログラムや文法，辞書の開発および保守に膨大なコストと時間が必要になるからである．

2.2 Web ディレクトリ拡張の手続き

図 1 に，Web ディレクトリ拡張手法の概要を示す．これは，各カテゴリに対して関連 Web サイトを発見する左側の部分と，関連 Web サイトに対して説明文を発見する右側の部分の 2 つに大きく分けられる．

Web ディレクトリの各カテゴリに対して人手で登録した少数の権威ある Web サイト集合と，Web ロボットを用いて自動的に収集した膨大な Web ページ集合の 2 種類を入力データに使用する．前者は，関

連 Web サイトを発見するための起点となる Web サイト集合であり，**シード Web サイト**と呼ぶ．後者からは，Web ページ間のリンク関係，各 Web ページ内のリンクの順序，リンクのアンカテキスト，およびソースアンカの直後のテキストなどの情報を抽出して，関連 Web サイトと説明文を発見するために使用する．

本システムは，Web ディレクトリの各カテゴリに対して，次の手続きを適用する．

(1) あるカテゴリのシード Web サイト集合に対して共参照関係にある Web サイトを発見する．
(2) 発見された Web サイトをスコア付けして，そのカテゴリに対して発見された Web サイトの中で高い共参照度を持つ Web サイトだけを**関連 Web サイト**として選択する．
(3) 関連 Web サイトをリンクしている Web ページ集合から，その Web サイトに付けられた紹介文を探す．
(4) 発見された紹介文をスコア付けして，最も高いスコアを持つ紹介文を説明文として選択する．

この結果，各カテゴリの内容に適合した関連 Web サイトとその説明文の組の集合を得ることができる．Web ディレクトリの各カテゴリでは，手動で管理されたシード Web サイトと自動的に発見された関連 Web サイトが混在し，被参照度順に表示される．

さらに，特に関連性が高いと思われる権威ある Web サイトを共参照度か人手による判断に基づいて選択して追加することで，さらにシード Web サイトを拡張することもできる．

3. 関連 Web サイトの発見

3.1 Web サイトの定義

Web ディレクトリでは Web ページ単位ではなく Web サイト単位で扱うので，Web ロボットで収集した Web ページ集合から Web サイト集合にまとめなおす必要がある．

Web サイトは，一般的には同一人物または同一団体によって記述された，1 つの情報源としての役割を果たす Web ページ集合と定義できる．しかし，さまざまな Web ページとハイパーリンクで接続されているために，Web サイトとしての境界が不明瞭であったり，大きな Web サイトがトピックや作成者が異なる複数の小さな Web サイトを含んでいるような包含構造も存在し，人間が直接見ても判断に困ることも多い．つまり，この定義に基づいて Web ページ集合から Web サイトに自動的にまとめなおすのは困難である．

そこで本論文では，この定義をさらに簡略化して，Web サイトをファイルシステム上の同一ディレクトリ内に存在する Web ページ集合と定義する．つまり，2 つの異なる Web ページがファイル名部分を除くと同一の URL プリフィックスを持つ場合に，同一の Web サイトに属すると見なす．これは，同一人物または団体によって作成される特定のトピックの Web ページは同一ディレクトリの中にまとめて置かれることが多いという経験則に基づいている．この定義では，複数のディレクトリにまたがる巨大な Web サイトが存在した場合にディレクトリ単位に断片化して処理される可能性があるが，共参照解析では利用者が Web サイトのトップページだと認識して多くリンクしている部分だけが発見されるので，最終的に得られる結果は最初に述べた一般的な定義とほぼ一致する．

3.2 共参照解析

同じ Web ページから，同時に 2 つの Web ページがリンクされている場合に，この 2 つの Web ページは共参照関係にあると呼ぶ．共参照解析は，文献の共引用分析[2] と同様に，共参照関係にある Web サイトの間には何らかの関係があると見なし，それを定量化することで Web ページ間の相関関係を求める手法である．共参照解析によって求められる共参照の度合いを，**共参照度（co-citation degree）**と呼ぶ．共参照度は，たとえば，与えられた Web ページに関連した Web ページ集合を発見し，順位付けするために使用される．

共参照解析アルゴリズムとしては，たとえば Dean らの提案した Co-citation アルゴリズム[3] がある．本論文ではアルゴリズムの起点として複数の Web ページを用いた場合の適合度を向上させた**Multi Co-citation アルゴリズム**を使用する．

なお，本論文では，共参照解析に Web サーバ間をまたがるハイパーリンクだけを使用する．この理由は，Web サーバ内のハイパーリンクは，たとえばナビゲーションメニューが機械的に生成されることも多いが，これに対して Web サーバ間をまたがるハイパーリンクは，第三者から見た Web サイトの信頼度および関連度の高さをよく表すからである．

ここで，共参照解析で Web サイトの関係を示す用語を定義しておく．Web サイト v の中のある Web ページが，別の Web サイト w の中のある Web ページにリンクしている場合には，v は w の**親（parent）**であり，w は v の**子（child）**であると呼ぶ．また，Web サイト w と Web サイト x が共通の親 v を持つ場合に，x は w の**兄弟（sibling）**であると呼ぶ．

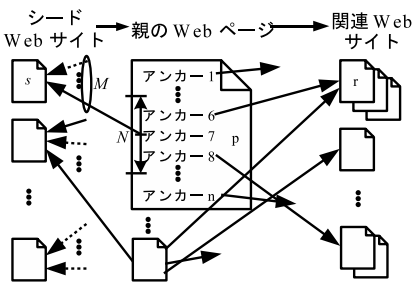

図 2 Web サイトの共参照関係
Fig. 2 The co-cited relation of Web sites.

3.3 Co-citation アルゴリズム

共参照関係にある Web サイトの関係を**図 2** に示す．

Co-citation アルゴリズムでは，まず起点となるシード Web サイト s に対して，s とリンク的に近い Web サイト群のグラフ構造である**近傍 Web グラフ（vicinity web graph）**を作成し，リンクを逆方向にたどって得られる親 p を無作為に最大 M 個まで選択して，s の親集合 P を作成する．次に，P の各 Web ページ p に対して，s のソースアンカ（リンク元のアンカ）周辺のリンクを近い順から最大 N 個まで選択してから，各リンクを順方向にたどって r を選択する．ソースアンカ周辺のリンクだけを使用するのは，同じ Web ページ内であっても，リンク間の距離が離れているほど，別の話題である可能性が高くなるからである．このときの s と r の関係が**共参照（co-citation）**であり，$p \Rightarrow (r, s)$ と表す．さらに，次のように Web サイト s に対する r の共参照度 $C(r)$ を計算する．

$$C(r) = |\{p | p \Rightarrow (r, s)\}| \quad (1)$$

ただし，Web ディレクトリに適用する場合には，各カテゴリに登録されている複数の Web サイトに関連する Web サイトを発見する必要があるが，オリジナルの Co-citation アルゴリズムでは単一の Web サイトしか扱えない．そこで Web ディレクトリ拡張では，次のように入力集合 S の各 Web サイト s がシード Web サイト集合の各 Web サイトに対してどれくらい高い共参照を示すかの総和を計算する．

$$C'(r) = \sum_{s \in S} |\{p | p \Rightarrow (r, s)\}| \quad (2)$$

3.4 Multi Co-citation アルゴリズム

上記アルゴリズムでは，たとえばシード Web サイト集合の中の 1 つだけに非常に高い共参照を示すが，他に対しては非常に低い Web サイトにも，高い共参照度が得られる．しかし，そのような Web サイトは必ずしもカテゴリ全体の内容とは適合しておらず，発見精度を低下させる要因になる．

そこで，次のように共参照度 $C_m(r)$ を定義し，本論文ではこれを**Multi Co-citation アルゴリズム**と呼ぶ．

$$C_m(r) = |\{s | p \Rightarrow (r, s) \wedge s \in S\}| + \alpha \times \sum_{s \in S} |\{p | p \Rightarrow (r, s)\}| \quad (3)$$

第 1 項は，ある Web サイトがシード Web サイト集合全体に対してどれくらい高い共参照を示すかどうかを表す項であり，すでに述べたカテゴリ内容とは別の理由でごく一部だけと高い共参照を示す場合の値は小さくなる．第 2 項は，$C'(r)$ に α を乗したものであり，実際には第 1 項を補正するために α に小さな値を指定して使用している．α の値が大きいほど，得られる結果は $C'(r)$ に近づくことになる．

3.5 トピックドリフト問題

共参照解析と同様に関連 Web サイトを見つけるリンク解析アルゴリズムとしては，Dean らが提案した Companion アルゴリズム[3] などがあげられる．このようなリンク解析に基づくアルゴリズムでは，被参照数が非常に多い Web サイトが，内容が適合しないのにもかかわらず結果として得られることがあり，これを**トピックドリフト問題**と呼ぶ[4]．

たとえば，近傍 Web グラフの back-and-forward 部分だけを使用して，より良い精度が得られるように改良したアルゴリズムである Companion-アルゴリズム[5] を実際に実装して比較したが，Co-citation アルゴリズムよりもトピックドリフトが顕著に発生する傾向が観測された．

これは，Companion-アルゴリズムでは収束するまで再帰的に計算を繰り返すために，局所的な関連性しか見ない Co-citation アルゴリズムよりもトピックを一般化する性質が強く，これは簡単には分からない潜在的な関係を発見する能力に優れている半面，本研究で対象とした Web ディレクトリのような細かいトピック分類のような応用には適していないと推測される．そこで，本論文では，Co-citation アルゴリズムを基にした．

しかし，Co-citation アルゴリズムも，シード Web サイト集合の中の 1 つだけに非常に高い共参照を示すような場合には，カテゴリの内容に適合しない Web サイトが得られてしまう問題が発生した．本論文で提案する Multi Co-citation アルゴリズムは，トピックドリフト問題をさらに改善する手法である．

3.6 Web ディレクトリ拡張への適用

Web ディレクトリの各カテゴリに登録されている Web サイト集合をシードとして，カテゴリごとに Co-

citation アルゴリズムまたは Multi Co-citation アルゴリズムで共参照関係にある Web サイトの共参照度を求める．

次に，各カテゴリに対して発見された Web サイトの中で指定された閾値より高い共参照度を持つ Web サイトを関連 Web サイトとして選択する．公開実験では，発見サイト数よりも発見精度を重視した結果，共参照度が 3.0 以上の Web サイトだけを使用している．

同一 Web サイトが複数のカテゴリに対して発見された場合には，公開実験に使用した Open Directory の登録方針に合わせて，一番高い共参照度を示すカテゴリだけに登録する．

4. Web サイト説明文の抽出

4.1 Web サイトの紹介文と説明文

Web ディレクトリの編集者にとって，各 Web サイトについての説明文を書くのは時間がかかる作業である．

一方，リンク集にある Web サイトが追加されるときには，同時にそのリンク先の Web サイトの紹介文が書かれる傾向がある．この紹介文は簡潔でありながら，概要をうまく表していることが多い．

そこで，本論文では，ある Web サイトにリンクしているリンク集の Web ページの中から，その Web サイトについて記述されている紹介文を抽出し，さらに得られた紹介文の集合の中から最も適切な文章を Web サイトの説明文として使用する．

4.2 Web サイトの紹介文の発見

Web サイトの紹介文を発見するために，実際のリンク集などで繰り返し使用される，Web ページの題名などが書かれたアンカの直後にその紹介文が続くパターンに注目する．そのようなパターンは HTML ファイル中では HTML 要素を用いて記述され，Web ブラウザでは**図 3** のように表示される．

実際の HTML ファイルは，必ずしも論理的にマークアップされているとは限らず，デザインを重視して意図した表示を実現するために，ときには HTML 要素を設計意図に反して使用していることが多い．このような場合，人間は視覚でアンカテキストと紹介文の組というパターンを容易に認識できても，文書の論理的構造を解析するだけでパターンを抽出することは難しい．しかし，そのようなデザインに重点が置かれた Web サイトが信頼できる情報源であることも少なくない．

そこで論理的な文書構造と視覚的な文書構造という点に着目すると，HTML 要素は次の 3 種類に分類できる．

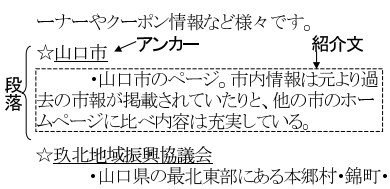

図 3 アンカとそれに続く紹介文のパターンの例
Fig. 3 A pattern example of an anchor and its following description.

(1) 文を論理的に開始，分割または終了させる要素（例，P，TABLE，UL，OL，DL，H1〜H6，DT，TR，LI など）
(2) 文を視覚的に分離する要素（例，BR，HR など）
(3) 文の分割に影響しない要素

(1) と (2) の要素はアンカテキストと紹介文の組のパターンの境界を表すが，前者が論理的にマークアップされた境界であるのに対して，後者は表示から推測される境界である．HTML ファイルを解析する時には，(1) と (2) の要素の性質を考慮しながら文の並びを表す中間表現に変換する．このときに，(3) の要素は無視する．この中間表現の中では，上記の論理的な文区切りを示す要素（開始要素，終了要素ともに）は 2 つの連続した改行に変換され，視覚的な文区切りを示す要素は単独の改行に変換される．そして，改行が 2 つ以上連続する箇所だけをパターンの区切りと見なし，さらにアンカテキストで始まる紹介文を抽出する．

もちろん，より良い内容の紹介文が存在しても，この方法で抽出できない可能性がある．しかし，より複雑なパターンの紹介文を抽出すると，抽出誤りも増加する恐れがあるので，本論文では精度の方を重視する．実際には，共参照解析で発見される関連 Web サイトは，その定義から被リンク数が多くなるので，多くの紹介文が存在することになる．そこで，ある良い紹介文が抽出できなくても，他の妥当な紹介文を抽出できる可能性は高く，すべての紹介文を網羅的に抽出する必要性は低いと考えられる．

4.3 良い紹介文の選択

一般的に，1 つの Web サイトには複数の紹介文が存在するが，そのすべてが Web ディレクトリの説明文として適切とは限らない．発見されたパターンのアンカテキスト部分，その直後に続く紹介文部分，それが掲載されている Web ページの 3 つに対して，説明文として適切だと思われる性質を次に示す．

- 適切なテキスト長．長すぎても，短すぎても，よい説明文とはいえない．

- 適切な文構造．一部が欠落していたり単なる単語の並びのような不完全な文，画像が含まれている文などは，説明文として適切とはいえない．
- 適切な用語．Web サイトの紹介に適した用語が使用されていなければならない．
- リンク集らしさ．一般の文章の中よりも，リンク集の中の紹介文の方が簡潔で客観的である．

また，一般的な Web ページではアンカが文の一部であるパターンがよく見かけられる．しかし，リンク集ではすべてを文章で説明するのではなく，Web ページの名前を示すアンカ部分とその紹介文の組からなるパターンの繰返しで表現することが多いので，たいていはアンカ部分と紹介文が分離されている．

そこで，これらの性質を考慮して，紹介文のスコア S を次のように定義し，それぞれの各紹介文のスコアを計算し，一番高いスコアを持つ紹介文を説明文として使用する．

$$S = S_t \times S_a \times S_r \quad (4)$$

S_t は，パターンの紹介文部分に関するスコアであり，文の長さ，句読点の数，特別な用語や記号の有無から導出される．文の長さが 50 字のときを，Web ディレクトリの説明文として表示するのに適切な長さと考えて，一番高く 100 と評価し，それより短いまたは長い場合は差の 2 倍を減点する．たとえば，Open Directory の日本語説明文は，平均文字数は 35.5 文字，最小文字数は 3 文字，最大文字数は 150 文字であるが，単なる単語の羅列で構成されたり，短すぎて十分に内容を説明しているとはいえない説明文も多く存在するので，本手法では，より長い紹介文を抽出する．さらに句読点が存在する場合には，種類ごとに 10〜20 の間の異なる重みを付けて，その出現数に応じて加点する．たとえば,「。」などの日本語特有の句読点の場合には高く評価し,「.」などの，日本語特有とはいえず，また句読点以外の用途に用いられそうな場合は低く評価する．逆に「☆」などの句読点以外の記号が出現する場合には出現数に応じて 10 の重みを付けて，その出現数に応じて減点する．また「公式」,「ホームページ」,「サイト」などの説明文に多用される用語が出現する場合にも，用語に応じて 20〜40 を加点する．

S_a は，パターンのアンカテキスト部分に関するスコアである．デフォルト値を 1 とし，アンカテキストが存在しない場合には 0.1 に，またはインライン画像が含まれている場合には 0.5 にスコアを減少させる．

S_r は，紹介文が存在する Web ページのリンク集らしさを示すスコアであり，抽出される紹介文数から導出される．同じ Web ページ内の他の紹介文数が 1 のときを 0.1 とし，紹介文数が 10 まで単調増加させ，10 以上の場合は 1 とする．

良い紹介文の例
アンカーテキスト：日本の酒
紹介文：日本酒造組合中央会の公式サイト。日本酒と焼酎、泡盛について歴史や製法、美味しい飲み方などを掲載している。日本酒や焼酎の違いについても解説。蔵元を検索したり、お酒に関する用語を調べることもできる。

悪い紹介文の例
アンカーテキスト：日本酒造組合中央会
紹介文：東京都港区西新橋 1—1—21 日本酒造会館　7FTEL 03–3501–0101

図 4 良い紹介文と悪い紹介文の例
Fig. 4 Examples of good and bad descriptions.

なお，これらの重みは現時点では経験に基づいて決定している．ここでは，複雑な自然言語処理は行わずに，文の外見的な特徴だけを使用していることに注意されたい．

説明文として良い紹介文と悪い紹介文の例を，**図 4** に示す．悪い例は単なる住所の紹介であり，リンク先を閲覧するかどうかの判断の役に立たない．

実際に，図 4 の紹介文に対してスコアを計算すれば，良い例では $S_t = 226$（97 字,「。」（20）が 4 個,「、」（20）が 3 個,「公式」（40）・「サイト」（40）の存在），$S_a = 1$，$S_r = 1.0$（紹介文数 23）で $S = 226$，悪い例では $S_t = 86$（43 字），$S_a = 1$，$S_r = 0.5$（紹介文数 5）で $S = 43$ となる．

5. 実　　験

5.1 Open Directory Project

Open Directory Project は，人手で編集されている最も巨大な Web ディレクトリのボランティアプロジェクトである．このプロジェクトには多くの編集者が参加しており，彼らが作成した膨大な量の Web カタログデータを一般公開し，AOL Search，Netscape Search，Google，Lycos，DirectHit，HotBot などの多くの有名な検索サービスで採用されている．

ただし，日本においては状況が異なる．Japanese カテゴリの下に登録されている Web サイトの数は日本の主要なポータルサイトと比較すると非常に少なく，多くの有名または重要な Web サイトが未登録である．これは，Japanese カテゴリを担当している編集者が少なく，さらに専任の編集者が存在しないカテゴリも多いからである．このために，日本の検索サービスでは Open Directory Project の Web カタログデータはほとんど採用されていない．

本論文では，Japanese カテゴリの下のカテゴリ集合

と，Web ロボットで収集した日本語の Web ページ集合に対して Web ディレクトリ拡張手法を適用し，その結果を関連 Web サイト発見と，良い説明文発見の 2 点について評価した．

実験用のデータを収集した 2000 年 12 月の時点では，Japanese カテゴリの下には 702 カテゴリが存在し，Web サイトの登録総数は 6,143 サイトであった．さらに，これらの Web サイトを起点として，JP ドメインに存在する Web ページおよび日本語を含むアンカテキストでリンクされている Web ページを 8 日で収集した．収集した Web ページ集合の規模は約 1,100 万ページであり，異なる Web サイト間のリンクは約 2,100 万本であった．Web ロボットの収集アルゴリズムの詳細は省くが，基本的には被リンク数が多い URL を優先することで，有益な Web サイトの情報を広くカバーすることを目指している．

5.2 関連 Web サイト発見の評価

関連 Web サイト発見の品質を評価するために，Co-citation アルゴリズムと Multi Co-citation アルゴリズムの結果を比較した．なお，経験に基づいて各定数値として $M = 2000$，$N = 10$，$\alpha = 0.1$ を使用している．N を抑え，α を小さな値にすることで，発見数よりも発見された Web サイトの内容の適合性を重視した設定にしている．

最初の実験では，Open Directory のカテゴリに登録されている Web サイトを本手法で自動分類しなおした場合について評価する．まず Open Directory の Japanese カテゴリの下の登録 Web サイト数が 4 つ以上のカテゴリを抽出する．これは 474 カテゴリであった．次に，抽出された各カテゴリから無作為に Web サイトを 1 つ選択し，評価用 Web サイト集合を作成する．評価用 Web サイトを除いた各カテゴリの Web サイト集合に 2 種類のアルゴリズムを適用して，評価用の各 Web サイトが元のカテゴリの共参照度が高い順で上位 10 件以内に分類されるかどうかを調べる．ここでは異なるアルゴリズムの比較を行うのが目的なので，閾値を指定して選択せずに，上位から同数の Web サイトを抽出して比較している．この分類精度 P を，次のように定義する．

$$P = \frac{D_r}{D_t} \tag{5}$$

D_t は，評価用 Web サイト集合の中でいずれかのカテゴリの上位 10 件に分類された Web サイト数である．D_r は，評価用 Web サイト集合の中で再び元のカテゴリの上位 10 件に分類された Web サイト数である．ただし，すでに述べたように，本手法では Web サイトを共参照度が一番高いカテゴリだけに分類するので，あるカテゴリに分類された Web サイトが別のカテゴリに対して，より高い共参照度を示すことはない．

この結果，Co-citation アルゴリズムは 0.80，Multi Co-citation アルゴリズムは 0.81 という結果が得られた．これは 700 以上のカテゴリが存在することを考慮すると，比較的良好な数値であると考えられる．また，元のカテゴリに分類されなくても，たとえば “ビジネス/食品/飲料/酒類” カテゴリと “レクリエーション/グルメとドリンク/酒類” カテゴリのように非常に類似した内容を持つカテゴリに分類されていることが多かった．

ただし，この実験では，基本的に権威ある Web サイトの分類精度を評価していることに注意されたい．つまり，リンク解析アルゴリズムでは，本質的にこのような被参照数が多い Web サイトが得られやすい傾向を持ち，さらに，このような権威ある Web サイトそのものの再分類にはトピックドリフト問題の影響は少ないために，本実験で得られる精度の差が出にくかったと考えられる．

2 番目の実験では，2 種類のアルゴリズムを用いて発見した関連 Web サイトの適合性を，被験者に評価させた．被験者は 20 代から 40 代の 8 名の男性であり，日常的にインターネットで情報を探している計算機科学またはコンピュータネットワークの研究者である．

まず，被験者に全カテゴリの中で特に熟知している 4 つのカテゴリを選択させた後で，互いに重複しないように 2 つのカテゴリを選択して，**表 1** に示すカテゴリを得た．次に，各カテゴリに 2 種類のアルゴリズムを適用して共参照度が高い順に Web サイトを 10 個選択し，その URL だけを並べたリストを作成した．なお，専任編集者が存在するカテゴリは，C，L，O，P だけである

被験者に，まずシード Web サイトのリストを見てカテゴリの内容を理解してから，実際に関連 Web サイトの内容を Web ブラウザで閲覧し，「関連していない（−2）」，「あまり関連していない（−1）」，「不可（0）」（アクセスできない，またはミラーサイトの場合），「関連している（+1）」，「非常に関連している（+2）」の 5 段階で評価してもらった．

この結果から，実験に使用した各カテゴリに対して平均適合度 R を計算した．**図 5** と**図 6** に，2 種類のアルゴリズムの各カテゴリに対する Web サイトの適合度を示す．なお x 軸の右に行くにつれてシード Web サイト集合が大きくなるようにカテゴリを配置している．

表 1 評価に使用したカテゴリ
Table 1 Categories used for our evaluation.

	カテゴリ名（サイズ）
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/イベント/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/グルメとドリンク/酒類/ワイン (10)
I	家庭/料理/食材 (10)
J	各種資料/辞書・辞典 (10)
K	社会/時事/自然災害 (11)
L	ビジネス/情報産業/通信/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/インターネット/WWW /ホームページ検索 (32)

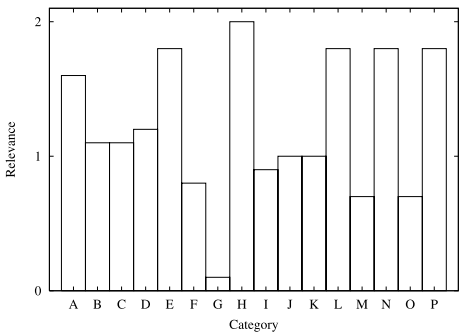

図 5 各カテゴリに対する適合度（Co-citation）
Fig. 5 The relevance for each category (Co-citation).

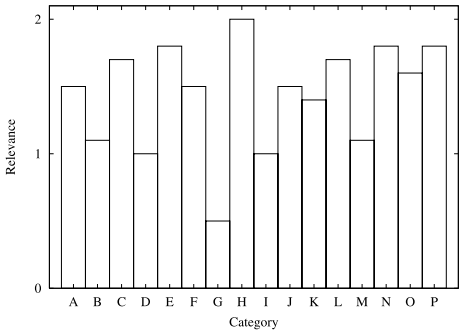

図 6 各カテゴリに対する適合度（Multi Co-citation）
Fig. 6 The relevance for each category (Multi Co-citation).

平均適合度と分散は Multi Co-citation アルゴリズムが 1.44 と 0.15，Co-citation アルゴリズムが 1.21 と 0.27 である．シード Web サイト集合の大きさと適合度の分布の相関関係は特に見られないが，Multi Co-citation アルゴリズムは，あまり良い結果が得られない G，M のようなカテゴリでも Co-citation アルゴリズムよりも良く，O のように著しい差がついているカテゴリも存在する．内容の適合度の点では，かなり良い結果が得られていることから，Multi Co-citation アルゴリズムではトピックドリフト問題がかなり改善されていることが分かる．

なお，Multi Co-citation アルゴリズムでも良い結果が得られなかったカテゴリでは，次の 2 つの原因が推測された．1 つは，専任編集者が存在しないカテゴリでは登録要求があった Web サイトだけが登録されるので，有名 Web サイトであっても登録されていないことが多く，近傍 Web グラフが小さくなりすぎることである．専任編集者が存在する C，L，O，P では安定して良い結果が得られているので，編集者が少しでも権威ある Web サイトを追加できれば改善できると推測する．もう 1 つは，リンク集のトピック分類が，Web ディレクトリのカテゴリ分類ほど詳細ではないことである．たとえば，洋画のカテゴリ G とアドベンチャーゲームのカテゴリ M が特に結果が悪い．これは，一般的なリンク集では洋画と邦画，アドベンチャーゲームとシミュレーションゲームのように細かい区別はせずに，映画，ゲームのような大まかな分類で済ますことが多いからだと考えられる．このような場合には，本手法で適合度を向上させることは難しい．

5.3 説明文発見の評価

8 名の被験者に，Open Directory の編集者が記述した説明文と，本手法で発見された説明文の適合性を評価してもらった．7 名は関連 Web サイト発見の適合度評価のときと同じである．

まず，表 1 のカテゴリから，サイト移転や廃止などの理由ですでにアクセスできなくなっている Web サイトを取り除いた後で，関連 Web サイト発見アルゴリズムおよび説明文発見アルゴリズムを適用し，105 個の Web サイト・編集者の説明文・抽出された説明文の組が得られた．

次に，これらを 8 つに分割し，各被験者に Web サイトの内容の真偽と適性（例，概要や特徴が表されているかなど），内容の重複（例，アンカテキストやカテゴリ名の繰返しなど），意味のない文字などの混入，客観性に注意するように指示を与えて，−2（不適合）から 2（高適合）の 5 段階で評価してもらった．

この結果から各適合度ごとに説明文の総数を求めて，**図 7** に示す．

平均適合度は編集者の説明文の 1.15 に対して，発見された説明文は 0.51 であり．0 以上の割合は編集者の説明文の 89.3%に対して，発見された説明文が

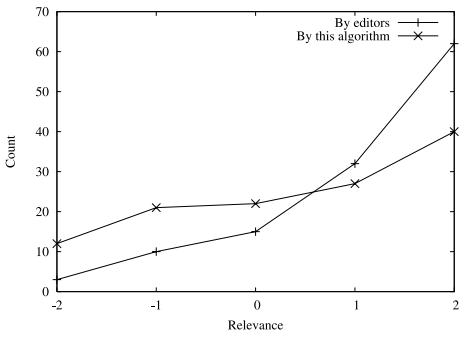

図 7 説明文の適合度ごとの総計
Fig. 7 The total of description's relevance.

73.0%であった．さらに，発見された説明文の方が高く評価された場合が 26（21.3%），同等に評価された場合が 36（29.5% ）存在し，編集者の説明文に及ばないにしても十分良い結果が得られている．

適合度が低くなる主な要因は，編集者の説明文の場合は，簡潔すぎる記述と内容の理解の誤りであり，発見された説明文の場合は，私的な意見や交友関係の記述などであった．なお，編集者の説明文は客観的であるが，発見された説明文は主観的なものも多かった．この実験では，客観性を優先するように指示を与えたために，結果的に発見された説明文の適合度が低くなる大きな要因となっている．しかし，Web サイトの評価や良い利用法などの主観的な記述は，実際にはリンク先の Web サイトを訪問するかどうかの決め手になる良い情報であり，内容が適切であれば有益であると考えられる．

6. 利用分析

6.1 ODIN ディレクトリ

ODIN ディレクトリは，Open Directory の Japanese カテゴリに本手法を適用して自動作成した Web ディレクトリである．Multi Co-citation アルゴリズムを用いて関連 Web サイトを発見し，ある閾値以上の共参照度を持つ Web サイトを，一番高い共参照度を示すカテゴリにだけ登録する．Web サイトを 1 つのカテゴリだけに登録する方針は Open Directory と同じである．

各カテゴリでは，権威あるサイトほど見やすくなるように，**図 8** のようにシード Web サイトと関連 Web サイトを区別せずに被リンク数順に並べるが，関連 Web サイトには “(*1)” のような引用記号を説明文の最後に付加して区別する．この記号は**図 9** に示すような Web ページの最下部にある引用 Web ページリストの各 URL を指す．明示的に引用元を示す理由は，引

図 8 カテゴリ内のシード Web サイトと関連 Web サイト
Fig. 8 Seed web sites and the related web sites in a category.

図 9 カテゴリ内の参考 Web ページリスト
Fig. 9 References in a category.

用元の Web ページの著作権を尊重していることに加えて，そのような Web ページは，そのカテゴリのトピックに関連したリンク集としても有用だからである．つまり，ODIN ディレクトリは，同時に Web ページの自動引用システムであるといえる．

なお，図 9 の下部に「このカテゴリにはエディタがいません」と表示されているが，このような専任編集者が存在しないカテゴリでは，Web サイトは自発的ではなく，外部からの要求に基づいて登録されるために，多くの重要な Web サイトが欠落しがちである．そのような場合であっても，ODIN ディレクトリでは被リンク数が多い権威ある Web サイトを補うことができる．

6.2 拡張された Web ディレクトリの有効性

ODIN ディレクトリは，2001 年 4 月から 2002 年 3 月まで一般公開実験を行った．実験期間中は不定期に Open Directory の Japanese カテゴリの下のカテゴリ集合を元にデータを作成した．最終更新の 2001 年 11 月 15 日の時点では，1,885 カテゴリ，20,428 サイト存在し，使用した Web ページ集合は約 1,300 万ページであった．共参照度 3.0 以上の関連 Web サイトを発見した結果，6,565 サイトが発見され，Web ディレクトリ全体に占める本手法で抽出された説明文を持つ関連 Web サイトの割合は 0.243 であった．

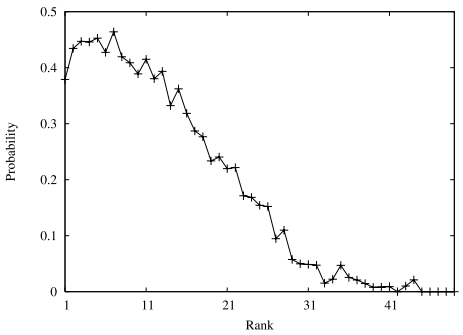

図 10 順位と関連 Web サイト率
Fig. 10 Ranks and probability of related Web sites.

まず，すべてのカテゴリに対して順位ごとにシード Web サイトか関連 Web サイトかどうかを集計し，ある順位の Web サイトが関連 Web サイトである平均確率を計算した．これを**図 10** に示す．カテゴリ内の Web サイトは被リンク数順であるので，順位は特定のトピックに関する権威度を表すと解釈できるが，実際に高い順位ほど有名な Web サイトが数多く補われていた．なお，このグラフでは 5 位がピークである．その理由は Open Directory のデータは，編集者が登録した有名 Web サイトと，依頼に基づいて登録された新規 Web サイトに大きく分類されるが，本手法はこの間を補っているから，つまり編集者がいないために登録されなかった，または編集者が見逃した有名 Web サイトを自動的に登録しているからと推測される．

次に，利用者が実際に閲覧した Web サイトについて解析した．ODIN ディレクトリでは，利用者の閲覧履歴を収集するために，各カテゴリの Web サイトのリンクには，その Web サイトの URL ではなく，閲覧履歴記録プログラムへの URL を指定している．そして，利用者の閲覧時には，そのプログラムがアクセス日時，カテゴリ ID，カテゴリの Web サイト数，閲覧 Web サイトの URL と順位などの情報を記録した後に，目的の URL に転送している．

たとえば，2001 年 12 月は，Web サイトは 137,278 回閲覧され，関連 Web サイトの閲覧率は 0.247 であり，これは全体に占める関連 Web サイトの率とほぼ等しいことから，シード Web サイトとほぼ同等の有用性を持っていると推測できる．

7. 関 連 研 究

7.1 共参照解析

Web ページの共参照関係を利用した研究は，他にもいくつか存在する．

Pitkow らは，Web ページの共参照関係に基づいた階層的クラスタリング手法を提案した[6]．この手法はある共参照頻度（co-citated frequency）以上の Web ページを解析しているだけなので，有用なクラスターだけでなく，多くの断片的なページ集合も抽出されている．

Dean らは，関連する Web ページを発見するために，Co-citation アルゴリズムと Companion アルゴリズムを提案し，それを比較した[3]．この Co-citation アルゴリズムでは，Pitkow らの場合と異なり Web ページ内の共参照される URL 間の距離も考慮されているが，単一の URL が対象である．

村田は，参照の共起性に基づいて近傍 Web グラフを作成し，その中でリンク元とリンク先の完全 2 部グラフを抽出して，コミュニティとして発見する手法を提案した[7]．この手法も共参照解析に基づくと考えられ，複数 URL が対象であるが，完全 2 部グラフではシード Web サイト数が大きい場合には，1 つも関連 Web サイトを発見できなくなる可能性がある．

大槻らは，経験則に基づいた自動生成した各地方公共団体の URL を参照し，かつアンカテキストが地域名であるリンク集（ハブ）を地方ごとに発見し，それが参照している Web ページ群を，それらの題名，アンカテキスト，強調テキストに編集者が作成したカテゴリ固有語辞書の用語が含まれているかどうかで詳細分類し，地域情報 Web ディレクトリを作成した[8]．この手法は，単純な共参照関係解析と，精度改善および詳細分類のための内容解析を組み合わせていると考えられる．ただし，特定の分野に依存しない Web ディレクトリの場合には，高い分類精度を実現できるような多くの辞書を作成する方法が明らかではなく，その作業コストも無視できない．

津田らは，Web ページのメタデータに基づいて分類することで，地域・ジャンル多観点自動ディレクトリを作成した[9]．この手法では，ある分野の Web ページ集合を探し出すブートストラップ過程で使用する関連度をリンク解析によって得られる参照度と共参照度を用いて求め，各カテゴリの選別と並べ替えで使用するページの人気度は PageRank を変更したものを用いているが，各カテゴリへの分類にはルールベースで抽出したメタデータを用いている．いったんメタデータを付与した後はさまざまな観点でカテゴリを見せることが容易であるが，適切なメタデータ抽出規則を作成する方法が明らかではなく，その作業コストも無視できない．

7.2 Web ページの自動分類

テキストに基づく自動分類は単語分布の類似性を利

用するので，同義語や多義語の存在により単語分布が異なる Web サイト間の内容の類似性の発見や，詳細に分類された Web ディレクトリのカテゴリに対する Web ページの分類，そしてその中からの権威ある Web ページの抽出は一般に困難である．このような分野には，ハイパーリンクと，それによって関連付けられた Web ページを解析する手法が有効である．

Chakrabarti らは，Kleinbreg の HITS アルゴリズムを元に，さらにソースアンカ周辺の一致する単語数を考慮して分類精度を改善する手法を提案した[1),10)]．

Glover らは，リンク元 Web ページのソースアンカ付近のテキストを利用して SVM で Web ページを分類した[11)]．

Toyoda らは，Companion アルゴリズムの精度を改良した Companion-アルゴリズムを用いて Web コミュニティを抽出した[5)]．ただし，Web コミュニティと Web ディレクトリの間には分類精度や分類観点の違いが存在し，必ずしも同一視できない[12)]．

7.3 テキストの自動要約

テキストの自動要約技術は自然言語処理の主要な研究テーマの 1 つであるが，文が統制されていない Web ページを，Web ディレクトリに使用される簡潔で短い説明文にまで要約できる技術は少ない．

Amitay らは，Web ページではアンカとそれに続く文で始まる段落が頻繁に使用され，それがリンク先の内容の要約であることに基づいた Web サイトの自動要約技術を提案した[13)]．

Radev らは，Web ページをクラスタリングし，クラスタに含まれる複数の Web ページから要約を作成し，文書を推薦するシステムを提案した[14)]．このシステムは Web サーチエンジンの検索結果を要約するために使用されるが，類似手法を Web ディレクトリのような静的な Web サイトに適用するのも興味深い．

8. おわりに

本論文では，Web ロボットで収集した膨大な Web データを使用して，Web ディレクトリを自動的に拡張する手法について述べた．さらに，発見された関連 Web サイトと説明文に対して被験者を用いた実験を行うとともに，一般公開されたシステムのログから利用者の行動を分析し，テキストの類似性を使用せずにハイパーリンク構造を基にした手法の有効性と可能性を示した．

本システムは限られた予算と人員で運営している小規模 Web ディレクトリの保守を妥当な品質で自動化できるだけでなく，十分な予算と人数で運営している巨大な Web ディレクトリにおいても，編集者が行う権威ある Web サイトを選択し，その適切な説明文を記述する作業を軽減することができる．また，主観的な紹介文を積極的に収集すれば，Web サイトの評価のような他の目的にも適用できると思われる．つまり，本システムは，ハイパーリンク解析を基にしたシステムの潜在的な可能性を示しているといえる．

なお，現在は，各カテゴリに対する関連 Web サイトの発見，発見された Web サイトに対する良い説明文の発見，および各カテゴリの Web サイトの順位付けの 3 種類のリンクベースの手法を互いに独立に使用しているが，これらを統合することで，より良い結果が得られる可能性がある．

謝辞 Open Directory のデータを配布している Netscape Communications Corporation と，Open Directory Project のボランティア編集者，および本実験に協力してくれた被験者に感謝する．

参考文献

1) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol.46, No.5, pp.604–632 (1999).
2) Small, H.: Co-citation in scientific literature: A new measure of the relationship between two documents, *In Journal of the American Society for Information Science*, pp.265–269 (1973).
3) Dean, J. and Henzinger, M.R.: Finding related pages in the World Wide Web, *Computer Networks* (*Amsterdam, Netherlands: 1999*), Vol.31, No.11–16, pp.1467–1479 (1999).
4) Bharat, K. and Henzinger, M.R.: Improved algorithms for topic distillation in a hyperlinked environment, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, Melbourne, AU, pp.104–111 (1998).
5) Toyoda, M. and Kitsuregawa, M.: Creating a Web Community Chart for Navigating Related Communities, *In Conference Proceedings of Hypertext 2001*, pp.103–112 (2001).
6) Pitkow, J. and Pirolli, P.: Life, Death, and Lawfulness on the Electronic Frontier, *Proceedings of the Conference on Human Factors in Computing Systems CHI'97* (1997).
7) 村田剛志：参照の共起性に基づく Web コミュニティの発見，人工知能学会論文誌，Vol.16, No.3, pp.322–329 (2001).
8) 大槻洋輔，佐藤理史：地域情報ウェブディレクトリの自動編集，情報処理学会論文誌，Vol.42, No.9, pp.2310–2318 (2001).
9) 津田　宏，鵜飼孝典，三末和男：Web ディレク

トリのためのページメタデータの自動付与の試み，情報学シンポジウム 2002, 情報処理学会情報学基礎研究会，pp.17–24 (2002).
10) Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P. and Rajagopalan, S.: Automatic resource compilation by analyzing hyperlink structure and associated text, *Proceedings of the 7th International World Wide Web Conference* (1998).
11) Glover, E.J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D.M. and Flake, G.W.: Using Web structure for classifying and describing Web pages, *Proceedings of the 2nd International World Wide Web Conference* (2002).
12) 吉田 聡, 豊田正史, 喜連川優：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, *DEWS2003*, 電子情報通信学会データ工学研究会 (2003).
13) Amitay, E. and Paris, C.: Automatically Summarising Web Sites - Is There A Way Around It?, *CIKM*, pp.173–179 (2000).
14) Radev, D., Fan, W. and Zhang, Z.: WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System, *Proceedings of NAACL Workshop on Automatic Summarization* (2001).

(平成 15 年 12 月 20 日受付)
(平成 16 年 4 月 7 日採録)

（担当編集委員 中野 美由紀）

風間 一洋（正会員）

昭和 63 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在 NTT 未来ねっと研究所主任研究員．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM 各会員．

原田 昌紀（正会員）

昭和 49 年生．平成 10 年東京大学大学院総合文化研究科広域科学専攻修士課程修了．同年日本電信電話（株）入社．情報検索の研究に従事．現在 NTT 未来ねっと研究所所属．

佐藤 進也（正会員）

昭和 38 年生．昭和 63 年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．協調作業における情報活用支援の研究に従事．現在 NTT 未来ねっと研究所主任研究員．電子情報通信学会，Internet Society，ACM 各会員．