Webページ集合からのサイト再構成の一手法

池田哲夫[†1] 森憲一[†2]
竹野浩[†3] 佐藤哲司[†4]

近年，Webマイニングや Webページの有用性評価等の分野において，Webにおける最小の情報単位をページではなく「サイト」とすることが有効であるとする研究が増えている．これらの研究におけるサイトは，その一般的な解釈である特定の個人あるいは組織が作成・管理するひとまとまりのページ群であるとされているが，この定義を満たすサイトを大量の Webページ集合から再構成する有効な手段はいまだ確立されていない．そこで，筆者らは，(1) 機械学習の一手法である決定木を用いてサイトのトップページとサイト間の境界とを決定すること，(2) 決定木作成の際に考慮する属性として，URL のパス表記およびハイパーリンク構造で与えられるページ間の関係と，ページ内に記述される特定表現の出現状況の 2 つを扱うこと，を特徴とするサイト再構成方式を提案する．さらに，提案法のプロトタイプを実装し，約 4,000 万の日本語が記述されている Webページ集合からサイトを再構成して，その有効性を検証した．

A Method for Reconstituting Web Sites

Tetsuo Ikeda,[†1] Ken'ichi Mori,[†2] Hiroshi Takeno[†3] and Tetsuji Satoh[†4]

Many recent research papers in the fields of web mining and information quality assessment use the term "web site" to refer to the minimum unit of information on the WWW. Unfortunately, there has, up to now, been no effective way of reconstituting web sites from a huge collection of web pages. We propose an effective solution that has two characteristics. One, it use the decision-tree algorithm, a machine-learning algorithm, to determine the root pages and site boundaries. Two, the relationships between pages and the presence of specific expressions in a page are considered when selecting the attributes required by the decision-tree algorithm. We introduce and test a prototype to verify the effectiveness of our method.

1. はじめに

近年，Webマイニングや Web情報の有用性評価等の分野において，WWW における最小の情報単位を Webページ（以下，単にページと呼ぶ）ではなく「サイト」とする研究が行われつつある[1),2),11)~13)]．これらの研究におけるサイトの解釈はほぼ一貫して「一般的に各個人，組織によって提供されていると認識される情報の単位」[13)]，つまり特定の個人あるいは組織が作成・管理するひとまとまりのページ群を意味している．

WWW からの知識発見において，ある情報が記述されたページが有用であるかどうかの最終判断は情報を吟味する各個人に依存する．この判断において重視される要因は，当該ページに記述された情報そのものの内容もさることながら，そのページが存在するサイトにおける情報量やサイトの作成者である個人あるいは組織に対する評価等も含まれる[1)] ため，サイト内の他のページを参照可能とすることでページの有用性判断を容易にできると期待される．また WWW の成長やリンク構造のマクロな理解のためには，WWW をサイトの集合ととらえ，サイト間のリンクに着目する必要がある[2)]．

これらの応用で重要となるサイトは，一般的な解釈で与えられるサイトで妥当と思われるが，大量のページ集合からサイトを再構成する有効な手段はいまだ確立されていない．先の研究例では，同一 Web サーバ

†1 岩手県立大学大学院ソフトウェア情報学研究科
Graduate School of Software and Information Science, Iwate Prefectural University
†2 日本電信電話株式会社 NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation
†3 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
†4 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

に含まれるページ群をサイトとする[1]，同一サーバの同一パス上に存在するページ群をサイトとする[12] 等の近似を行っている．

サイトの定義は上記のとおり非形式的なものであり，大量のページ集合からサイトを再構成する厳密な手法を導くことは困難である．そこで，筆者らは以下を特徴とする，上記の一般的な解釈に沿ったサイトの再構成方式を提案する．(1) 機械学習の一手法である決定木を用いてサイトのトップページとサイト間の境界とを決定する．ここで，サイトはトップページを根とする木構造として構成することとする．(2) 決定木作成の際に考慮する属性として，URL のパス表記およびハイパーリンク構造で与えられるページ間の関係と，ページ内に記述される特定表現の出現状況の 2 つを扱う．

以下，本論文の構成について述べる．2 章では筆者らが提案するサイトの再構成方式を説明し，3 章では約 4,000 万ページの日本語が記述されているページ集合に提案方式を適用した結果を述べる．4 章では関連研究との比較を行い，5 章で考察を述べる．

2. サイト再構成方式

サイトの一般的な定義「特定の個人あるいは組織が作成・管理するひとまとまりのページ群」は非形式的なものであり，サイトを構成する要素が厳密に規定されているわけではない．そこで筆者らは，サイトのトップページの決定あるいはサイト間の境界の決定に役立つと思われる属性をリストアップし，それらの属性を基に機械学習機構により決定木を作成することで，上記定義を満たすサイトを再構成する．

以下，提案方式を下記ステップに分けて説明する．

- サイトのトップページの決定
- サイトの間の境界の決定（サイトの内部構造の決定）

2.1 サイトのトップページの決定

サイトのトップページの決定は，まずページをトップページとそれ以外のページに分類する決定木を作成し，それを利用してページ集合からトップページを選別する．以下，決定木の作成方法と，決定木を利用してのトップページ決定方法を説明する．

2.1.1 決定木の作成方法

決定木の作成は，ページを分類するために使用する属性の決定フェーズと，与えられた属性に基づいて分類の学習を行う学習フェーズからなる．

(1) 属性の決定フェーズ

まず，既存の研究[5] において有効性が示されている，リンク構造とディレクトリ構造に関する下記の属性を採用することとした．

- ファイル名が index.html，top.html，welcome.html 等である．
- URL のパス部分のファイル名と，そのファイルが存在するディレクトリ名とが一致している．
- フレームページである．
- バックリンクを含む．
- URL のパス中に「home」「top」「welcome」等の文字列を含む．
- 他のページからのバックリンクがある．
- サーバ内部ページへのリンクの本数．
- サーバ外部ページへのリンクの本数．

ただし，バックリンクとは「戻る」等のアンカ文字列を持つリンク，もしくはそのページが存在するディレクトリの先祖ディレクトリに存在するページへのリンクである．

次に属性としてページ内に記述されている特定表現の出現状況を考慮する．どのような特定表現がトップページの判別に有効であるかを以下の実験により求めた．約 4,000 万のページ集合からランダムに選択した 6,912 ページを被験者に見せ，それがサイトのトップページであるかを判定させた．トップページであると判断した場合，トップページであると判断する際に用いた属性を上位 3 位まで記録する．実験の結果，トップページの判別に有用な属性として以下が抽出された．

(1) タイトルに「ホームページ」等の表現を含む．
(2) カウンタ CGI を含む．
(3) メールアドレス表記を含む．
(4) 「ようこそ」，welcome 等の表現を含む．
(5) 「このサイトは」「このページは」等，自サイト（ページ）の説明語句を含む．
(6) 「更新日」等更新説明語句を含む．
(7) 「copyright」等著作権関連表記を含む．

なお,「〜等」とある部分に関しては，それぞれ類似の 3〜5 パターンの文字列（ワイルドカードを用いた正規表現を含む）に対するマッチングをとることが有効であることを意味する．

ここで抽出された属性の再現率と適合率を**表 1** に示す．再現率は，全トップページ（549 ページ）のうち各属性を含むページの割合であり，適合率は，全ページ（6,912 ページ）のうち各属性を含むページのうち，トップページの割合を示す．表より，単独の属性の再現率，適合率はあまり良くない．また，精度良くトップページとその他を分類できる属性の組合せを求めることは困難といえる．そこで，機械学習により決定木

表 1 各規則の再現率と適合率
Table 1 Recall and precision of each rule.

項目	該当件数	再現率	適合率
a	289	0.11	0.5
b	174	0.49	0.78
c	124	0.54	0.31
d	93	0.24	0.35
e	90	0.24	0.35
f	86	0.63	0.84
g	77	0.36	0.41

を作成・利用することとした．

（2）**学習フェーズ**

先に述べた既存研究において有効とされている属性および上記実験で抽出した属性を用いて，分類の学習を行った．学習データは前記の 6,912 ページである．

学習には事例分類に適した決定木アルゴリズム C4.5[9] を用いた．C4.5 は，属性情報のエントロピーを用いたヒューリスティックス関数を最大化するように属性情報を再帰的に選択して決定木を構成し，学習データへの過剰適応を避けるため，信頼レベルに応じた枝刈りを行うアルゴリズムである．属性情報がとりうる値は，{a,b,c} 等の列挙型の値と，整数等の連続値の双方が適用可能である．本学習フェーズでは，上記属性のうち，真偽を問う属性に対しては 0/1 の値を割り当て，リンクの本数を問う属性に対しては連続値を割り当てた．

学習の有効性を検証するため，交差検定（cross validation）を行った．検定にあたり，学習事例 6,912 ページをランダムに 5 つのページ集合に等分し，4 つの集合で学習した決定木を残りの 1 つの集合に適用するプロセスを 5 回実行した．この結果，新規事例に対する平均予測誤り率は 14%であり，交差検定の値は 86%であった．これより，学習の精度はほぼ十分と考えられる．

2.1.2 トップページの決定方法

得られた決定木を用いて，以下のステップでトップページを決定する．まず，準備としてトップページ尤度を定義する．トップページ尤度とは，あるページのトップページとしての尤もらしさであり，C4.5 の出力における判別の確信度によって与えられる値である．

ある同一 Web サーバ上に存在するページ集合ごとに，以下のステップを繰り返してトップページを決定する．

ステップ 1：データの整列

ページ集合に含まれる各ページのトップページ尤度を決定木を用いて導く．次いで，それらのページを，URL のパスの深さの昇順，トップページ尤度の降順，URL 文字列の昇順にソートしてページリストに格納する．トップページリストを空にする．

ステップ 2：トップページの決定

ページリストからページを取り出し，下記の規則に基づきトップページか否かを決定し，トップページならば，トップページリストに加える．

(a) トップページを決定するに際して考慮すべきことは，インターネットサービスプロバイダ（ISP）が提供するサーバに複数のユーザがページを持っていることである．これらのページはユーザごとにサイト化されることが望ましい．提案法ではディレクトリの特定の階層（具体的には第 1 階層あるいは第 2 階層）に存在するページのトップページ尤度の平均値が閾値以上である場合，その階層に属する各ディレクトリ中で最大のトップページ尤度を持つページはすべてトップページであるとする．閾値は以下に記す方法に基づき 0.35 とした．

(b) (a) にあてはまらないページに関しては，トップページ尤度が閾値以上である場合，トップページとする．閾値は以下に記す方法に基づき 0.5 とした．

閾値を決めた方法について説明する．

(a) の閾値については，約 100 のプロバイダの提供するユーザホームページスペースを持つサーバを対象に，ユーザサイトのトップページが存在するディレクトリ階層のトップページ尤度の平均値を調査し，すべての調査対象ディレクトリにおける最小値 0.35 を閾値とした．

(b) の閾値については，C4.5 のデフォルト値である 0.5 を用いて学習したところ十分高い精度が得られたので，その値を用いることとした．

なお，各サーバ内の全サイトが，最初に選択されたトップページ集合の中のいずれかからリンクで必ずたどれるとは限らない．そのような非連結サイトが存在する場合は，新たにトップページを選択する．選択方法は次節で述べる．

2.2 サイト間の境界の決定（サイトの内部構造の決定）

サイト間の境界の決定は，各ページが属するサイトを決定することで達成される．サイトを構成するページは，その代表であるトップページからリンクをたどって到達可能であると考えられる．ここで問題となるのは，あるページに複数のトップページから到達可能であった場合，どのトップページからのリンクをそのページへの適切なリンクとすべきかということであ

る．筆者らは，この問題を解決するため，Mizuuchi らの提案するエントランスパス（entrance path）[8] 概念を用いた．ページの作成者は，ページの読み手が同一のパスをたどることを仮定して，サイトの設計を行うことが多い．このパスがエントランスパスである．

各ページへのエントランスパスを正しく決定することにより，ページを適切なトップページに結びつけることができると期待される．この考えに基づき，本論文に示すサイトは，トップページからたどれるエントランスパスによって形成される木構造をなすページ集合である．ここで，トップページに結びつけるエントランスパスを構成する個々のリンクを「幹リンク」と呼ぶこととする．

サイト間の境界を発見することは 2 つのステップからなる．まず，各リンクに対して幹リンクとなる尤もらしさ（幹リンク尤度と呼ぶ）を決定する．次いで，決定された幹リンク尤度を用いて，サイト間の境界を決定する．

2.2.1 幹リンク尤度の決定方法

幹リンク尤度とは，C4.5 の出力における判別の確信度によって与えられる値である．幹リンク尤度を決定するアルゴリズムは，エントランスパスを発見する Mizuuchi らのアルゴリズムに基づいているが，2 つの点で異なる．1 つ目は，幹リンクを決定するための情報として，リンク構造に加えて，アンカ文字列やターゲット属性値等の情報を用いている点である．2 つ目は，より精度の良いアルゴリズムを導くために機械学習を用いている点である．

まず，幹リンク尤度を得るための機械学習に用いた属性情報を説明する．

（1）**リンクパターン**

Mizuuchi ら[8] が述べているように，リンクパターン，すなわちリンクの元ページと先ページのディレクトリ上での関係のパターンは，そのリンクが幹リンクとなるか否かと強い相関があるため，属性情報とした．

ここで採用したリンクパターンの種類を以下に示す．

- 同一ディレクトリに存在するページへのリンク
- 直接下位ディレクトリに存在するページへのリンク
- 間接下位ディレクトリに存在するページへのリンク
- 兄弟ディレクトリに存在するページのリンク
- 直接上位ディレクトリに存在するページへのリンク
- 間接上位ディレクトリに存在するページへのリンク
- 別ドメインに存在するページリンク
- フレームのソースページから構成ページへのリンク

（2）**リンク元ページとリンク先ページのディレクトリツリーにおける距離**

ディレクトリツリーを枝の重みのない無向グラフと見た際のノード間の距離が大きくなるにつれ，それぞれのノードに対応するディレクトリに含まれるページ間の関係は希薄になると考えられる．リンク元，リンク先ページが存在するディレクトリ間の，ディレクトリツリーにおける距離（連続値）を属性情報とした．

（3）**バックリンク**

バックリンクは，その性質上多くの場合幹リンクになりえないと考えられる．あるリンクがバックリンクになるか否かの判別に必要な属性として，アンカ文字列に「戻る」等の特定表現が出現するか否かを採用した．

（3）**リンクタグの target 属性**

ウェブブラウザで表示した際，フレーム外へのリンクや，新しいウィンドウを開くリンク等は，リンク元ページとの関係が希薄であると考えられる．この種のリンクは，それが定義されているリンクタグの target 属性の値が “_top” や “_blank” となっている．そこで，属性情報として，リンクタグの target 属性が，1) 設定されている，2) 設定されており，かつその値が “_top” または “_blank” である，3) 設定されていない，を採用した．

次に，学習データの作成方法を説明する．約 4,000 万ページ間のリンクの中からランダムに 24503 リンクを抽出し，各リンクが幹リンクか否かを判別し学習データとした．

このデータを基に機械学習を行って決定木を作成し，交差検定による有効性の検証を行った．検定にあたり，学習事例 24,503 リンクをランダムに 5 つのリンク集合に等分し，4 つの集合で学習した決定木を残りの 1 つの集合に適用するプロセスを 5 回実行した．この結果，新規事例に対する平均予測誤り率は 16%であり，交差検定の値は 83%であった．これより，学習の精度はほぼ十分と考えられる．

2.2.2 サイト間の境界の決定方法

サイト間の境界は以下の 3 ステップで決定する．

ステップ 1：トップページの取り出し

トップページ決定法で求めたトップページリストから 1 ページずつ取り出してステップ 2 へ．トップページリストが空で，かつ，すべてのページがいずれかのサイトに割り当てられたときは，ステップ 3 の終了処

理へ．そうでない場合は，未処理ページの内，以下の条件を満たすページをトップページとしてステップ 2 へ．

- ディレクトリ階層が最も浅いページ集合に属する．
- トップページ尤度が最も高い．

ステップ 2：トップページを根とする木の構成

トップページからリンクをたどって到達可能なページの集合とその親子関係を求める．一般に，あるページには複数の入りリンクが存在することから，あるページへの入りリンクがすでに他のトップページからの幹リンクとして割り当てられている場合や，現在割当てを行っているトップページから複数の入りリンクでページが指示される場合がある．このような場合には，どの入りリンクを幹リンクとして割り当てるかを選択しなければならない（注：ページを唯一のトップページと結びつけるためには，ページへの入りリンクの 1 つだけを幹リンクとする必要がある）．幹リンクの選択法を**図 1** を用いて説明する．

この図は，ページ n3 に対してトップページ t1 からページ n1 を経由して幹リンクが割り当てられている状態で，ページ n3 の別の入りリンクがトップページ t2 からページ n2 を経由して幹リンクを割り当てようとしている場合を示している．この場合に，ページ n2 から n3 への幹リンク尤度がページ n1 から n3 への値より大きければ，ページ n3 はトップページ t1 からの木から切り離してトップページ t2 の配下に移し替える．このような操作が許容される条件は以下の 3 項目である．

- 親となるページ（幹リンクの始点となるページ）を変更した，すなわち，幹リンクを付け替えることで，閉路ができないこと．
- 新たに設定しようとする幹リンクの幹リンク尤度が，元の値より大きい．
- 新たに設定しようとする幹リンクの幹リンク尤度は元の値と変わらないが，トップページからの平均幹リンク尤度が大きくなる．

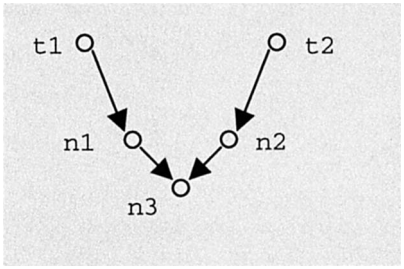

図 1 幹リンクの選択の例
Fig. 1 Example of selection of a stem link.

ここで，平均幹リンク尤度とは，トップページから当該ページへのリンク列（図の例では t1→n1→n3 等）のそれぞれの幹リンク尤度の平均値である．このような平均値を用いるとした理由は，リンクをたどって幹リンクを割り当てる処理において，それまでにたどったリンクの幹リンク尤度の合計値とリンクの本数を保持しておけば，簡単に計算できるからである．

ステップ 3：終了処理

トップページを根とし木構造をなすページ集合をサイトと定義し，サイトを識別するためのサイト ID，トップページ，サイトを構成するページ集合を，サイトの属性としてデータベースに格納する．

3. 評　　価

提案したサイト構成機能（トップページ決定機能とサイト間境界決定機能）のプロトタイプを C，Ruby を用いて試作し，日本語が記述されている 37,149,045 ページからサイトを再構成した．処理を実行したマシンの仕様は，CPU が PentiumIII 1.2 GHz，メモリが 2 GB であり，このマシンを 5 台並列に動作させた．

3.1 サイトサイズの分布

まず，サイトの構成ページ数（以下サイトサイズと呼ぶ）ごとの分布を測定した．測定結果を**図 2** に示す．測定結果から，構成ページ数の小さいサイトが多いことが分かった．

サイトサイズ 5 未満のサイトが全サイト数に占める割合を**表 2** に示す．

サイトサイズ 5 未満の約 300 サイトをランダムに選択し，その内容を調査したところ，約 30%は，個人あるいは小規模な組織が作成したサイト等，サイトと見なして問題ないページ集合であった．残りのサイトは，そのほとんどが**表 3** のいずれかであると想定されるものであった．

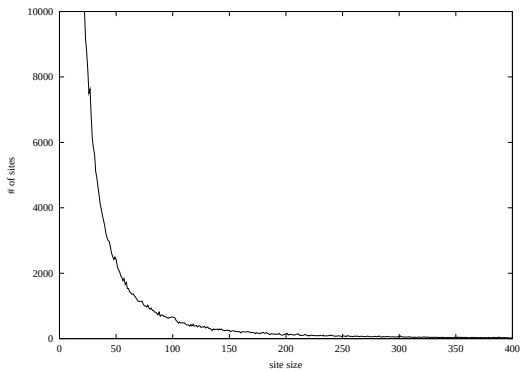

図 2 サイトの分布
Fig. 2 Distribution of sites.

表 2 サイトサイズ分布
Table 2 Distribution of site size.

サイズ	サイト数	全サイト比 (%)	全ページ比 (%)
1	6,654,903	81.3	17.9
2	405,366	5.0	2.2
3	191,579	2.3	1.5
4	120,313	1.5	1.3
5 以上	808,987	9.9	77.1

表 3 サイズの小さいサイトの分類
Table 3 Classification of small sites.

項目	割合 (%)
1. 他サイトへのリンクのみのページ	54.3
2. 移転サイトの残存ページ，製作中サイト等	9.0
3. リンクの切れた古いコンテンツ	6.0

他サイトへのリンクのみのページ（1.）のほとんどはサイト内にリンク以外の情報はなかった．移転サイトの残存ページやリンクの切れた古いコンテンツ（2., 3.）に該当するページは，URL を直接指定することでアクセス可能であるが，サイト作成者が意図的にサーバに残した可能性は無視できる程度に低いと判断して問題ないと考える．このようなページが少なからず存在する理由としては，サイトの更新の際に積極的にゴミ掃除（Garbage Collection）をせずに，新たに更新・作成したページをアップロードするだけで済ませるケースが多々あるためと考えられる．

3.2 サイト構成の評価

提案法により再構成されるサイトが「特定の個人あるいは組織が作成・管理するひとまとまりのページ群」という定義に沿っているかについて，以下の 2 つの観点から評価した．

- サイトのトップページの尤もらしさ．
- サイト間の境界の尤もらしさ．

評価に際しては，サイズの小さいサイトの多くは他サイトへのリンクのみ等であることから，サイズ 5 未満を除く 808,987 サイトを用いることとした．

次に，評価方法について説明する．まず，サイト集合から無作為に 400 サイトを抽出し，評価者がそのトップページが以下のいずれに分類されるかを確認することで，トップページが適切であるものとそうでないものを峻別する．

(1) サイトのトップページである．
(2) ある組織の内部組織のトップページであり，トップページと判断できる．たとえば，企業や大学等の大きな組織において，部，課等の単位で情報発信している場合である．
(3) サイトのトップページではない．

トップページの定義に対する妥当性に関しては，上記 (1)，(2) に相当するトップページの割合を計算し，確認する．また，(3) に該当するトップページに対して，失敗したケースの原因分析を行う．サイト間の境界の尤もらしさに関しては，(2) に相当するトップページを，評価者がその境界を精査して定義に沿った境界が構成されているものとそうでないものに峻別する．もし不適切な境界が存在する場合は，それらの事例に対して，失敗したケースの原因分析を行う．

評価対象を 400 サイトとした根拠を説明する．母集団において自然なサイトがどのように分布しているかが不明なため，自然なトップページの割合の検定はノンパラメトリック検定となる．現実的に許容できる誤差 5%でノンパラメトリック検定を行う場合は，十分に大きい母集団で約 400 の標本数で十分なことが知られている[10]．このため，400 サイトを標本として抽出した．

上記実験において，サイトのトップページを決定できていた割合は 83.2%であった．次に，(3) に分類されたトップページについて，それらがどのようなページであるかを調査・分析した結果を以下に示す．

- 上位ディレクトリ階層のページと同一の作成者によるコンテンツのインデクスページ（91%）：
  これには，ある企業サイト（http://www.a.com/）の「製品紹介」のインデクスページ（http://www.a.com/product/index.html）等が相当する．提案法では,「製品紹介」のページはその上位ディレクトリ階層のサイトと同一の作成者により記述されていることがその外観上明らか（たとえばデザインが同じ等）であっても，正しく判別されない場合がある．これはページのローカルな特徴量に強く依存したトップページの判別を行っている機械学習の限界，およびトップページの決定におけるステップ 2 の規則 (a) の弊害と考えられる．
- メニューが flash や javascript で記述されており，リンクがたどれない（5%）：
  提案法ではトップページからリンクされているページのみをサイトに帰属させているため，flash や javascript（特にソースがページ外部にあるもの）についてはリンクをたどることができず，同一サイトと判定されない．
- トップページが外部サーバに存在するページ（4%）：
  提案法では同一サーバ内のページをサイトにグループ化するため，トップページとコンテンツが

別のサーバに存在する場合はそれぞれが別のサイトと判定される．

なおこの調査において，もともとページが持っている曖昧さ，たとえば「会社等大きな組織においては，会社全体をサイトと見なすのか，部をサイトと見なすのかが明確でない」に起因するケースが確認された．ページが曖昧さを持つことは，誰もが情報発信することを許されているWWWの本質的な特徴であり，このような問題への対処は今後の課題である．

続いて，サイト間境界の判定については先に述べたように(2)に相当するトップページに対応するサイトに関して被験者を変えて評価した．その結果，すべての検査対象において判定境界が一致したことから，提案法が妥当なサイト境界を規定できるといえる．

4. 関連研究

Amento ら[1]はWebページ評価におけるサイト概念の有用性について言及し，大量収集したWebページ集合からサイトを再構成する方法を記している．同文献では，サイトは単一の個人または組織によって維持管理されている特定のトピックを持つページ集合であり，ドメインとは異なるものとする．その導出方法では，収集したページ集合に対して1. ISPサーバ等Webサーバごとのルール，2. 一般ルール（同一ディレクトリ階層に複数ページが存在する場合のトップページの選び方等）を適用し，トップページを得る．しかしこの方法では，ISPのサーバのように特殊なサイト決定規則がある場合はそのサーバごとにルールを書く必要があり，またその維持管理コストも無視できないと思われる．筆者らの実験結果から，このようなルールが発見されたWebサーバは全体の8%と無視できない割合を占めるため，サイトの再構成においてはこれらを精度良く発見可能な機構が求められる．

またサイトを情報単位としたWebマイニングの研究[2]があるが，ここではサイトはドメインと同一視されている．

一方，Webページの検索結果リストの閲覧性向上を主眼とした大量Webページの組織化技術が研究されてきた．Tajima ら[3]はWebページ集合を個別のトピックに対応する連結サブグラフに分割する手法を提案している．同手法は，リンク先ページの内容をベクトル空間モデルで比較し，類似度の高いページをグループ化する．類似したアプローチによる永藤らの成果[4]によれば，この手法では多くのページがグループ化されず，ページ数1のグループが約9割を占める．

ページの作成者は，閲覧者が同一のパスをたどることを仮定してサイトの設計を行うことが多い．Mizuuchi ら[8]は，それらのパス（エントランスパス）を発見する手法および同手法を利用したサイトの入口ページを見つける手法を提案している．同文献によれば，入口ページの発見についてはうまくいかない場合が多く，失敗するケースの考察においてディレクトリ構造の情報のみを利用する同手法の限界であると述べている．

Li ら[5]は，「論理ドメイン」概念を提案し，論理ドメインへの入口ページと論理ドメイン間の境界を識別する方法を提案している．論理ドメインは特定の意味的関連を持つページのグループである．この手法では論理ドメインを得るために各サーバにつき初期論理ドメイン数，最小論理ドメインサイズ，リンク半径（入口ページからたどるリンク数）の3つのパラメータを設定する必要がある．前節の実験結果から，Webサーバごとのサイトの数やサイトに含まれるページ数にはかなりのばらつきがあるため，Webサーバごとに論理ドメインとしてサイトを取り出すためにはすべてのWebサーバについて共通のパラメータを利用するのではなく，各Webサーバについて適切なパラメータを求める必要があると思われる．

風間ら[6]は，ディレクトリ構造に着目してWebページを組織化したページグループを提案している．この方法で得られるページグループとはおおむね同一ディレクトリに含まれるページ群であり，提案法により再構成されたサイトよりは小さなページ集合であると考えられる．

高野ら[7]は，他のWebサーバからリンクされているページを代表ページとするインフォメーション・ユニットを提案している．同文献にはインフォメーション・ユニットはサイトによるクラスタリングを実現し，その結果も妥当なものであると述べられている．同文献では定量的な評価データが示されていないため，提案法との実質的な比較は難しい．

5. 考察

5.1 小さいサイトの扱いについて

サイトの分析結果から，サイズが5未満のサイトのうち30%は意味のあるサイトであることが分かった．これは，全ページ数の1/15に相当する．

これらの小さいサイトの扱いは応用依存であるといえる．たとえばサイトの再構成結果をWeb検索エンジンに利用する場合，検索結果の網羅性が多少落ちても適合率を重視するポリシーを採用するならば，サイズの小さなサイトは捨てるという選択が考えられる．

逆に，再現率を重視するポリシーを採用するならば，サイズの小さなサイトをそのまま残す，あるいは近隣のサイトに統合する（たとえば，古いニュースやログからなるサイトの場合は，それらが以前に属していたと推定されるサイトに統合する）といった解が考えられる．

5.2 サイト構成の適切さについて

トップページの判別誤りの多くは提案法の限界を示唆し，また適合率向上のために導入した,「トップページの尤度の平均値が閾値を超えた場合，その階層の各ディレクトリには必ずトップページがあるとする」ルールに判別誤りの弊害があることが分かった．典型的なパターンとしては，尤度の平均値を計算する対象のディレクトリ数が少ない場合に，ごく少数の尤度の高いトップページ（これらはサイトと判断するにふさわしい）の影響でそれ以外のディレクトリもサイト化されてしまう場合があげられる．この場合については，当該条件適用対象の階層にあるディレクトリ数によって閾値を変化させる等の方法で，判別誤りをある程度軽減できると考えている．

また今回はサーバをまたぐようなサイトが別々のサイトと判定されてしまう問題もあるが，この点に関しても外部リンクの幹リンク尤度を適切に学習させることで，ある程度精度を向上させることが可能であると考えられる．ただし処理速度を多少犠牲にする必要があり，どちらを重視するかは適用先の要件に依存する．

5.3 処理時間について

プロトタイプでは，約 4,000 万ページを数日で処理することができた．この結果は 5 台のマシンを用いての値であるが，提案法はスケーラビリティが高く，並列処理で処理時間を短縮できると考えられる．すなわちトップページ決定の主要な処理は，ページ解析と決定木による判別であり，いずれもページ単位で処理することができる．またサイト間の境界の決定では，同一の Web サーバに含まれるページから各サイトが再構成されることから，Web サーバ単位で負荷配分を行うことができる．このため，いずれの処理もデータ分散によって高い並列度を達成できると考えられる．

5.4 適 用 分 野

サイト概念を導入することで，Web コンテンツの有用性評価において，より正確な判断を可能にすることが期待できる[1)]．また，関連サイト発見，サイト間の関連の可視化といったサイト概念の応用[11)~13)] 分野においても，提案法を用いることでより一般的な解釈に近いサイトを情報単位としたより精度の高い分析が可能になると思われる．

一方，サイトを情報単位とすることにより，ページ単位では得られなかったサイト単位での単語の出現頻度分布やページ間の親子関係，サイト内の指定キーワードを含むページ数等の情報を得ることができる．既存の WWW 検索システムのランキングにおけるページの評価基準，すなわちページおよびアンカテキスト内の指定単語の出現頻度情報やリンク構造に加えて，これらの情報を考慮することのできるサイト検索システムは，従来のページ検索を改善，あるいは補完することが期待される．

6. おわりに

大量のページ集合から，一般的な解釈に沿ったサイトを再構成する一方式として，(1) 機械学習の一手法である決定木を用いてサイトのトップページとサイト間の境界とを決定すること，(2) 決定木作成の際に考慮する属性として，URL のパス表記およびハイパーリンク構造で与えられるページ間の関係と，ページ内に記述される特定表現の出現状況の 2 つを扱うこと，を特徴とするサイト再構成方式を提案した．

また提案法を実装し，37,149,045 ページの日本語が記述されているページ集合からサイトを再構成し，80%以上の精度で一般的な解釈に沿ったサイトを構成しうることを確認した．今後は，サイト再構成のさらなる精度向上と，提案したサイト再構成方式を利用したサイト検索システムの実現が課題である．

参 考 文 献

1) Amento, H.B., Terveen, L. and Hill, W.: Does “authority” mean quality? Predicting expert quality ratings of Web documents, *23rd ACM SIGIR*, pp.296–303 (2000).
2) Bharat, K., Chang, B.-W., Henzinger, R.M. and Ruhl, M.: Who Links to Whom: Mining Linkage between Web Sites, *ICDM 2001*, pp.51–58 (2001).
3) Tajima, K., Mizuuchi, Y., Kitagawa, M. and Tanaka, K.: Cut as a Querying Unit for WWW, Netnews, e-mail, *Hypertext 1998*, pp.235–244 (1998).
4) 永藤拓宏，遠山元道：ページ群への分割を利用した WWW 検索エンジン，第 9 回データ工学ワークショップ (1998).
5) Li, W.-S., Kolak, O., Vu, Q. and Takano, H.: Defining logical domains in a web site, *Hypertext 2000*, pp.123–132 (2000).
6) 風間一洋，原田昌紀，佐藤進也：サーチエンジンの検索結果のマルチレベル・グルーピングの評価，コンピュータソフトウェア，Vol.17, No.4,

pp.58–69 (2000).
7) 高野 元，久保信也：サイテーション・エンジン：リンク解析を用いた WWW 検索ランキングシステム，情報処理学会データベースシステム研究会報告，No.120-2, pp.11–16 (2000).
8) Mizuuchi, Y. and Tajima, K.: Finding Context Paths for Web Pages, *Hypertext 1999*, pp.13–22 (1999).
9) Quinlan, J.R.: *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers (1993).
10) 東京大学教養学部統計学教室：人文・社会学の統計学，東京大学出版会 (1994).
11) Amento, B., Terveen, L., Hill, W. and Hix, D.: TopicShop: Enhanced support for evaluating and organizing collections of Web sites, *13th ACM UIST*, pp.201–209 (2000).
12) 原田昌紀，風間一洋，佐藤進也：参照共起分析の Web ディレクトリへの適用，情報処理学会情報学基礎研究会報告，No.61-8, pp.45–52 (2001).
13) 外山大介，吉高淳夫，平川正人：リンク構造の連結性に注目したコミュニティ導出に基づく Web ブラウジング手法の提案，情報処理学会情報学基礎研究会報告，No.61-8, pp.53–58 (2001).

(平成 14 年 7 月 8 日受付)
(平成 14 年 11 月 5 日採録)

池田 哲夫（正会員）

1979 年東京大学理学部情報科学科卒業．1981 年東京大学大学院理学系研究科情報科学専攻修士課程修了．同年日本電信電話公社（現 NTT）電気通信研究所入所．NTT 在籍中，データベース管理システム，データベース応用システムの研究開発等に従事．2002 年より岩手県立大学教授．専門は，データベース工学，情報検索等．博士（工学）．ACM，IEEE CS 各会員．

森 憲一

NTT サイバースペース研究所員．1998 年広島大学大学院工学研究科情報工学専攻修士課程修了．同年日本電信電話株式会社入社．以来，情報検索およびその応用技術の研究開発に従事．

竹野 浩（正会員）

NTT サイバーソリューション研究所主任研究員．1987 年大阪大学大学院基礎工学研究科修士課程修了．同年日本電信電話株式会社入社．以来，通信処理，情報検索の研究開発に従事．電子情報通信学会会員．

佐藤 哲司（正会員）

NTT コミュニケーション科学基礎研究所社会情報研究部主幹研究員・グループリーダ．社会インタラクション，情報検索，分散並列処理，マルチメディアデータベースに興味を持つ．現在，大阪大学大学院情報科学科客員教授．博士（工学）．電子情報通信学会会員．