論　文

Webサイト評価基準の抽出と評価の自動化

非会員　李　　鵬*　　非会員　山田　誠二**
非会員　新田　克己*

Web Site Evaluation Criteria Extraction and Evaluation Automation

Peng Li*, Non-member, Seiji Yamada**, Non-member, Katumi Nitta*, Non-member

This paper proposes an automated web site evaluation approach using machine learning to extract evaluation criteria from the existing evaluation data. Evaluating web sites is a significant task because evaluated web sites provide useful information for users to estimate sites' validation and popularity. Although many practical approaches have been taken to present a measuring stick for web sites, their evaluation criteria are set up manually. Thus, we develop a method to obtain evaluation criteria automatically and rank web sites with the learned classifier. Evaluation criteria are discriminant functions learned from a set of ranking information and evaluation features collected automatically by web robots. We conducted experiments and confirmed the effectiveness of our approach and its potential in performing high quality web site evaluation.

キーワード：評価基準抽出，Webサイト評価
Keywords：Evaluation Criteria Extraction, Web Site Evaluation

1. まえがき

世界のWeb利用者はすでに15億人を越え，巨大な情報ネットワークであるインターネットは急速に普及しており，我々は身近に便利で莫大な量の情報資源を手に入れている。こうしたインターネット上の情報資源として最も利用されているものの一つがWebサイトである。Webサイトは，豊富な情報量と高い更新性を備えている。しかし一方では，情報が不正確なサイトやスパムサイト，悪意のあるサイトも急速に増えている。検索を行う際，ユーザは情報の関連性だけではなく，その有用性も求めている[1]。したがって，Webサイトの質に関する評価が重要になっている。

現在，Webサイトの評価は企業，大学，国家機関，Eコーマスなどさまざまな分野で行われており，IBMやNTTデータなど，企業や個人向けにWebサイト評価を行うサービスを提供する会社も増えてきている。一方では，Webサイトの評価は，全幅の信頼を置くことのできる標準的な評価基準の設定は非常に難しい。評価基準はユーザによって異なり，違う領域のサイトに同じ評価基準を適用しても良い結果が得られる可能性は低い。さらに，同じサイトに対しても，時間が経てば評価が変わるケースも多い。ここ数年Webサイトのユーザビリティ，アクセシビリティ，セキュリティなど各方面の評価ツール[2]-[5]が多く開発されているが，評価基準の設定や評価データの収集は基本的には手動で行われる。しかし，手動でWebサイト評価を行うには処理できるWebサイトの規模に明らかに限界があり，コンピュータシステムによる自動的なWebサイトの評価が強く望まれる。Webサイト評価が自動化されれば，膨大な量のWebサイトを迅速に評価することが可能となり，さらには質の高いWebサイトの検索，推薦，収集そして情報の提供といった新たなサービスの市場を開拓することが期待できる。

このようなWebサイト評価の特徴を踏まえ，本研究では，既存の評価データからWebサイトの評価基準を自動抽出し，未評価のサイトに対して，抽出された評価基準による自動評価を実現する。具体的には，まず，Webサイトの評価をWebサイトのランク付けとして定式化する。次に，ユーザによってランク付けされた評価データを訓練データとして，機械学習を適用することで，ランキングの評価関数を獲得する。そして，獲得された評価関数を基に，任意の未評価サイトをランク付けし，Webサイト評価の自動化を実現する。

抽出された評価基準の妥当性と更新性を保つためには，二つの課題を解決しなければならない。一つは，いかに不特定多数のユーザの評価データを定期的かつ自動的に入手するかである。ランキングの学習は大量の評価データを必

* 東京工業大学大学院総合理工学研究科
〒226-8503　神奈川県横浜市緑区長津田町 4259
CISS, IGSSE, Tokyo Institute of Technology.,
4259 Nagatsuta, Midori,Yokohama, Kanagawa 226-8503
** 国立情報学研究所，総合研究大学院大学
〒101-8430　東京都千代田区一ツ橋 2-1-2
National Institute of Informatics SOKENDAI.,
2-1-2 Hitotsubashi,Chiyoda, Tokyo 101-8430

要とする。評価データは通常，人手による適合評価という形で得られる[6]が，本研究は登録型ランキングサイトに着目する。登録型ランキングサイトとは，ユーザ推奨と投票による有力サイトをリアルタイムにランキングするWebサービスで，ここ一年で著しい発展を遂げている。SiteRank† のような人気サイトには，数十万のサイトが登録されており，ユーザによるカテゴリ作成，サイト登録や投票が頻繁に行われている。我々は，この登録型ランキングサイトのカテゴリ化されたランキング情報を機械学習の訓練データとして利用し，カテゴリに依存したWebサイト評価を実現する。つまり，Webサイト評価の自動化を，ランキング情報を訓練データとして判別関数を学習する分類学習の問題としてとらえる。我々はRanking SVM[7]，Prank[8]，マルチクラスSVM[9]の三つの手法の比較実験を行い，この問題に適した学習方法を選択する。

もう一つの課題は，いかに妥当な評価属性を定期的かつ自動的に入手するかである。本研究はアクセシビリティやユーザビリティなど特定の側面からの評価ではなく，あるカテゴリにおけるWebサイトの総合的な評価を対象としている。そこでWebサイトの特徴を反映すると同時に，Webロボットによる定期的収集が可能なものを評価属性として選択する。そして，これらの評価属性を用いて抽出された評価基準の妥当性を評価実験によって検証する。

以降，2章では関連研究について述べ，本研究の位置づけを明確にする。3章では本研究が提案するWebサイト評価基準の抽出方法を説明する。4章では，登録型ランキングサイトの評価データを基に行ったWebサイト評価の実験について述べる。5章で考察を行い，6章で本論文をまとめる。

2. 関連研究

Webサイト評価は大きく分けて，自動評価と非自動評価の二つの方法がある。非自動評価には二通りのアプローチがある。一つは被験者実験で，ユーザにWebサイトを評価してもらい，評価基準に関する意見を収集し，分析する。もう一つのアプローチは，専門家に評価基準を指定してもらう。評価基準の作成に関する研究は数多く行われている。Alexander[10]らのWeb WISDOMやArone[11]らのWebMACなどは，情報源評価に関する専門領域である図書館情報学における研究結果から，資料の質に対する評価可能性が支持されてきた指標をメタ情報として利用し，評定基準を作成する方法を用いている。鷲見[12]らのWEIは，知りたい情報を得るために利用されるWebサイトの質を評価するための基準を提示した。しかし，多くの場合，どのアプローチにも実用上の課題が残されている。第一の理由は，Webテクノロジーの急速な発展により，Webサイトが非常に複雑化していており，人手による評価自体が困難になっていることである。第二の理由は，人手による評価は時間がかかる一方で，Webサイトの更新が早く，最新状態のWebサイトに対する評価を得るのが難しいことである。このため，Webサイトの自動評価が強く望まれている。

Webサイトの自動評価ツールは，サイトのユーザビリティ，アクセシビリティ，セキュリティなどの側面に焦点をあてたものが多く開発されている。それらは，次の五つのカテゴリに分けられる。1. サーバ性能の分析，2. 利用状況の分析（ログデータなどによる），3. WACGとSection 508のガイドライン準拠に関するチェック，4. ナビゲーションテキストの分析，5. 仮想的なユーザ（エージェントソフトウェア）によるナビゲーションシミュレーション[13]である。これらの自動評価ツールは，あらかじめ評価基準を設定し，自動収集可能な評価属性を用いることにより，ユーザが介入することなくWebサイトの評価を行うことができる。しかし，一般的には，Webサイトの評価を行う際の評価基準があらかじめ設けられていない。このため評価の過程が自動化されていても，評価基準を人手によって設定する必要があり，自動化のメリットが半減する。その理由の一つは，非自動評価と同じように，Webサイトは非常に複雑化していて，人手による評価基準の作成が困難になっていることである。もう一つは，評価基準自体が時間とともに変動し，常に新しい評価基準が求められることである。したがって，Webサイト評価の完全な自動化を実現するためには，評価基準の自動設定と，その評価基準に基づく自動評価の両方を実現する必要がある。

これまでも，学習によって評価基準を設定する研究はいくつかある。Velayathan[14]らはユーザのブラウジング行動からWebページの自動評価を行うアプローチを提案している。この研究では，ユーザが実際に閲覧したWebページに，ユーザの手により“興味ある”，“興味ない”のラベルを付けてもらい，それをクラスのラベルとする。そしてWebページのブラウジングパターンを評価属性とし，分類学習を行うことで，評価基準を自動的に獲得している。しかしながら，この研究で学習される判別関数は，Webサイトではなく Webページに対する評価であり，また評価の過程では，ユーザが実際に評価対象をブラウジングする必要があるため，自動評価を実現していない。一方，Richardson[1]らは独自の評価属性を採用し，RankNetアルゴリズムを用いて評価基準の抽出を実現している。しかしその評価属性は自動的に取得可能なもので構成されていないため，Webサイトを自動的に評価することができない。

Webサイトの完全な自動評価の試みとして，本研究では，評価基準の自動抽出と，それに基づいた自動評価を実現する。

3. 評価基準の抽出

本研究では，既存の評価データからWebサイトの評価基準の自動抽出を実現する。前述のように，抽出された評価基準の妥当性と更新性を保つためには，特定多数のユーザの評価データを定期的入手する必要があり，本研究では登録型ランキングサイトを評価データのソースとする。

† http://siterank.org/jp/

登録型ランキングサイトでは，カテゴリごとに分類された Web サイトが，ユーザの投票によりランク付けされている。このランキングは，多くのユーザが種々な視点から，対象となる Web サイトを評価した結果を投票し，集計したものであり，多くのユーザがもつ Web サイトの評価関数に基づいてランク付けされた結果である。したがって，ランキング情報を訓練データとして，その評価関数を機械学習により自動的に獲得し，この評価関数を用いて，未評価の Web サイトを自動的にかつ的確に評価することが可能となると期待できる。

このように本研究では，Web サイト評価の自動化を，ランキング情報を訓練データとして判別関数を学習する分類学習の問題としてとらえる。本章では，まずこのランキング学習問題を解決するいくつかの学習手法について述べ，続いてデータの属性，Web サイトの自動評価の手続きについて説明する。

〈3・1〉 ランキング学習手法 本研究はランキング情報を入力とし，ランキング情報出力とする学習問題に当たる。このような問題を解く学習手法として，Ranking SVM, Prank, マルチクラス SVM などがある。本節ではそれぞれの学習手法を説明する。

〈3・1・1〉 Ranking SVM Ranking SVM[7]は全順序関係のあるマルチクラスの判別関数を高精度で学習可能な学習アルゴリズムである。図 1 は，Ranking SVM におけるランキングメカニズムを表している。データ x から SVM の超平面 w までの距離を 1 次元空間にマッピングすると，$\theta_1,\ldots,\theta_{K-1}$ は距離の比較対象となる閾値にあたる。データ x のランクが i のとき，$\theta_{i-1} < w^T x_j < \theta_i$ が成り立つ。逆に，$w^T x_j$ の値がどの区間に入るかによって,そのランクが決まる。

まず，$\alpha_i = \theta_{i+1} - \theta_i$；$1 \le i \le K-1$（$K$ はランクの数）を定義する。ランクが 1 の場合 $w^T x_j$ は α_1 の左の領域，ランクが K の場合 $w^T x_j$ は α_{k-2} の右の領域に入る。また，ランク $i>1$ のとき，(1)式が成り立つ。

$$w^T x_j > \theta_{i-1}; w^T x_j + \alpha_{i-1} > \theta_i; w^T x_j + \sum_{k=i-1}^{K-2} \alpha_k > \theta_{K-1} \quad \cdots\cdots(1)$$

同様に，ランク $i < K$ のとき，(2)式が成り立つ。

$$w^T x_j < \theta_i; w^T x_j + \alpha_i < \theta_{i+1}; w^T x_j + \sum_{k=i}^{K-2} \alpha_k < \theta_{K-1} \quad \cdots\cdots(2)$$

Ranking SVM は超平面 w を $\overline{w} = [w, \alpha_1, \alpha_{2,\ldots}, \alpha_{K-2}]$ に拡張することによって，ランキング問題を従来の分類問題に置き換えることができる。n 次元の x_j を $n+K-2$ 次元に拡張し，

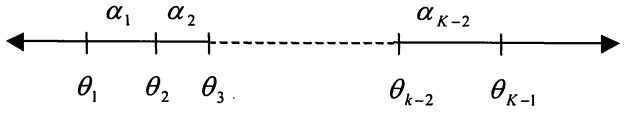

Fig. 1. Ranking mechanism.

$\overline{x_j^+}$ と $\overline{x_j^-}$ を以下のように定義する。

$$\overline{x_j^+}[l] = \begin{cases} x_j[l] & 1 \le l \le n \\ 0 & n < l < n+i-1 \\ 1 & n+i-1 \le l \le n+K-2 \end{cases};$$

$$\overline{x_j^-}[l] = \begin{cases} x_j[l] & 1 \le l \le n \\ 0 & n < l < n+i \\ 1 & n+i \le l \le n+K-2 \end{cases} \quad \cdots\cdots(3)$$

簡単な解析により，(4)式が成り立つことが分かる。

$$\overline{w}^T \overline{x_j^+} > 0 \Rightarrow w^T x_j > \theta_{i-1}; \overline{w}^T \overline{x_j^-} < 0 \Rightarrow w^T x_j < \theta_i \quad \cdots\cdots(4)$$

以上の定式化により，ランキング問題は，$\overline{w}^T \overline{x_j^+} > 0$；$\overline{w}^T \overline{x_j^-} < 0$ を満たす $\overline{w}$ を $n+K-2$ 次元で学習する問題に帰着される。

〈3・1・2〉 Prank Prank[8]はパーセプトロンによるオンライン学習アルゴリズムである。データ x□□^{n} からランク y□$\mathcal{Y}=\{1,2,\ldots,k\}$ へのマッピングルールを H:□$^n \to \mathcal{Y}$ とする。Prank は重みベクトル w□□^{n} と閾値のセット $b_1 \le \ldots \le b_{k-1} \le b_k = \infty$ を持つパーセプトロンモデルを学習によって得られる。新しい事象 x が与えられると，w と x の内積を計算し，ランクの予測値を $w \cdot x < b_r$ を満足する最小の閾値 b_r のインデックスと定義する。つまり，事象 x のランク予測値は

$$H(x) = \min_{r \in \{1,\ldots,k\}} \{r : w \cdot x - b_r < 0\} \quad \cdots\cdots(5)$$

となる。図 2 に Prank のアルゴリズムを示す。

図 3 の例を用いて Prank の学習過程を説明する。ランクは 5 段階あり，実際のランクは y = 4 である。したがって $w \cdot x$ の値は第 4 区間の b_3 と b_4 の間に入るはずである。しかしこの例では $w \cdot x$ は b_1 の左側に入り，予測ランクは 1 になっている。閾値 b_1 ,b_2 と b_3 は $w \cdot x$ よりも高くなってしまっていて，この間違いを修正するため，それぞれを b_1-1 ,b_2-1 と b_3-1 に置き換える。同時に w を $w+3x$ に修正し，$w \cdot x$ が 3□x□^{2} 大きくなる。この変化は図 3 の中央の図になる。そして右の図は修正後の状態である。

〈3・1・3〉 マルチクラス SVM マルチクラス SVM[9]は SVM を多値識別に応用した学習アルゴリズムである。

Initialize: Set $w^1 = 0, b_1^1, \ldots, b_{k-1}^1 = 0, b_k^1 = \infty$.
Loop: For $t = 1,2,\ldots,T$
- Get a new rank-value $x^t \in \mathbb{R}^n$.
- Predict $\hat{y}^t = \min_{r \in \{1,\ldots,k\}} \{r : w^t \cdot x^t - b_r^t < 0\}$.
- Get a new label y^t.
- If $\hat{y}^t \ne y^t$ update w^t (otherwise set $w^{t+1} = w^t, \forall r : b_r^{t+1} = b_r^t$):
  1. For $r = 1,\ldots,k-1$: If $y^t \le r$ Then $y_r^t = -1$ Else $y_r^t = 1$.
  2. For $r = 1,\ldots,k-1$: If $(w^t \cdot x^t - b_r^t) y_r^t \le 0$ Then $\tau_r^t = y_r^t$ Else $\tau_r^t = 0$
  3. Update $w^{t+1} \leftarrow w^t + (\sum_r \tau_r^t) x^t$.
     For $r = 1,\ldots,k-1$ update: $b_r^{t+1} \leftarrow b_r^t - \tau_r^t$

Output: $H(x) = \min_{r \in \{1,\ldots,k\}} \{r : w^{T+1} \cdot x - b_r^{T+1} < 0\}$.

Fig. 2. The Prank algorithm.

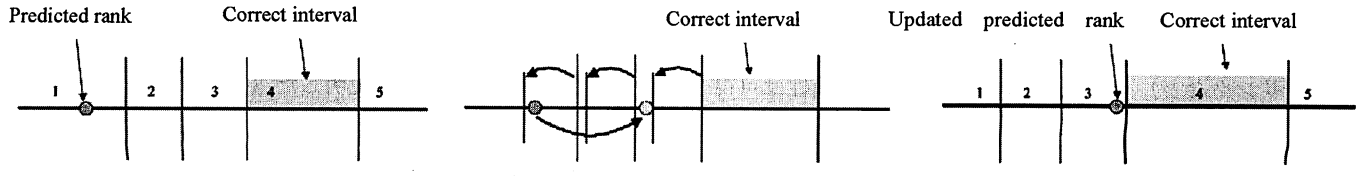

Fig. 3. An Illustration of the update rule.

SVM は基本的には 2 つのクラスを識別する識別器を構成するための学習法で，多クラスの識別器を構成するためには，複数の SVM を組み合わせるなどの工夫が必要となる。例えば，k クラスの分類問題を解くため，1 クラスとその他の残りのクラスとを識別する 2 クラス分類 SVM の集合 $f^1,\ldots, f^k$ を生成する。そして，符合関数を適用する前に，(6)式のような最大出力に従ってマルチクラス分類を行い，2 クラス分類 SVM の集合 $f^1,\ldots,f^k$ を統合しておく。

$$\arg\max_{j=1,\ldots k} g^j(x) \quad (6)$$

ここで，$g^j(x)$ は次式で表現される。

$$g^j(x)=\sum_{i=1}^{l} y_i \alpha_i^j \cdot k(x,x_j)+b^j \quad (7)$$

最終的に符合関数を適用する際は，以下のようになる。

$$f^j(x)=\mathrm{sgn}(g^j(x)) \quad (8)$$

〈3・2〉 評価属性の選択 本研究では，登録型ランキングサイトの Web サイトランキング情報を訓練データとする。ここで扱うデータは，複数の評価属性とその属性値で記述され，訓練データは，ランキングサイトで決められたランクを正しいラベルとしてもつ。一方，学習終了後に獲得された判別関数により，まだ評価されていない Web サイトをランキングすることができるが，そのような Web サイトを試験データとする。妥当な判別関数，つまり評価基準，を抽出するためには，適切な評価属性を選択することが重要である。

Web サイト評価には利用者とサイト作成者の二つの視点がある。利用者の視点から見ると，ほとんどの研究は Web サイトのデザインやコンテンツに着目している。Olsina[15]らは，Web サイトの質の評価の主な基準は機能，ユーザビリティ，効率と信頼性であると提案した。Huizingh[16]は Web サイトアーキテクチャをデザインとコンテンツに分類し，Web サイトの特徴にしたがってそれぞれの分野を評価基準の作成に使用した。Mateos[17]らは WAI(Web Assessment Index)モデルを開発し，スペインの大学のサイトを対象に実験を行った。Palmer[18]らは Web サイトのユーザビリティ，デザインとパフォーマンスの評価基準を設定し，被験者実験を行った。そして，Web サイトの優劣はスピード，ナビゲーション，コンテンツ，レスポンスと相互作用に依存するという結論に至った。

サイト作成者の視点から見ると，Web サイト評価に関する研究は Web サイトのユーザビリティとアクセシビリティに着目している。Sinha[19]らと Ivory[20]らは高質な Web サイトの共通の特徴を特定するため，専門家とユーザが推薦する Web サイトを調べた。Sinha[19]らは Webby Award 2000 データセットを対象に，高質なサイトとそうでないサイトの要素を識別した。文書，ストラクチャーとナビゲーション，ビジュアルデザイン，機能，相互作用の五つの分野の評価属性を評価基準の設定に使用した。実験の結果，文書は画像よりも重要であるが，評価基準はそれぞれ独立に考えることはできないという結論に至った。Ivory[20]らは Webby Award 2000 データセットと 157 個の Web ページを対象に同様の実験を行い，84%の高精度で分類が可能なことを確認した。

Table 1. 10 fields of evaluation features.

Field	The number of features
Top page's global link popularity	1
Freshness	2
Indexable text information	1
Multi-media contents	4
Accuracy of spelling and grammar	1
Accuracy of HTML documents	4
Contents security	5
Contents constitution	4
Design	5
Others	4

これらの研究を踏まえて，我々は利用者視点とサイト作成者視点の両方を考慮に入れ，評価属性を選択した。また，従来のアンケート方式による評価では，評価属性は主観的なものでも構わないが，本研究では評価の自動化を目的としているため，評価属性自体も客観的かつ自動的に取得できる必要がある。よって，本研究では，表 1 に示すような客観的に決定でき，かつ自動収集可能な 10 分野 31 個の評価属性を設定した。以下に，それぞれの評価属性の性質と計算方法について述べる。

(i) 広域のリンク重要度：ある Web ページの重要度を，リンクの構造のみから客観的に計算する広域のリンク重要度として Google が公開している PageRank を用いる。

(ii) サイトの鮮度：サイトの鮮度は，新しい情報と古い情報の情報量のコントラストを意味し，サイトの更新頻度と，更新されたテキストの情報量が全体に占める割合の二つの指標を用いる。

(iii) インデクシング可能なテキストの量：HTMLで書かれているWebサイトは可視的な部分とそうでない部分（ハイパーリンクなど）で構成されている[21]。この属性は可視的なHTMLテキストの量で測る。具体的には，HTMLタグを除外したHTMLソースファイルのサイズを計算する。

(iv) マルチメディアコンテンツの量：画像ファイルの数，動画ファイルの数，音声ファイルの数，フラッシュの数の四つの指標を用いる。

(v) スペルと文法の正確さ：現段階ではスペルチェックを行い，ミススペリングの数を測っている。

(vi) HTMLドキュメントの正確さ：これは，文字コードや画像サイズなどを指定しているか否かを用いる。

(vii) コンテンツの安全性：キャッシュ，スクリプト，Webロボットなどのコントロールに関連した項目をチェックしている。

(viii) コンテンツの構成：テキスト情報量に対する各マルチメディアコンテンツの割合を指標とする。

(ix) デザイン関連：BGMや背景画像の有無，フレームやスタイルシートの使用などの情報を採り入れている。

(x) その他：トップページリンク数，サイト全体のリンク数，作者情報やページに対する説明文を指定しているか否かを用いる。

現段階では上記31個の属性を扱っているが，これらはWebサーバからどのような情報を獲得できるかという技術的な制約を受ける。よって，今後Webの発展に伴うWebロボット，Webサーバの改善によって，より多くの評価属性を取得できる可能性はある。なお，上記の属性を用いて抽出される評価基準の妥当性は〈4・4〉節の抽出精度の評価実験で確認される。

〈3・3〉 Webサイト評価の自動化 本研究で提案するWebサイトの自動評価システムの概要を図4に示す。学習の段階では，登録型ランキングサイトから得たカテゴリ別のランキング情報とWebロボットで収集した各サイトの評価属性を学習し，カテゴリごとに評価基準となる判別関数を生成する。評価の段階では，評価対象のWebサイトの評価属性値を取得し，対応したカテゴリの判別関数に照らし合わせて予測ランキングを計算し，提示する。

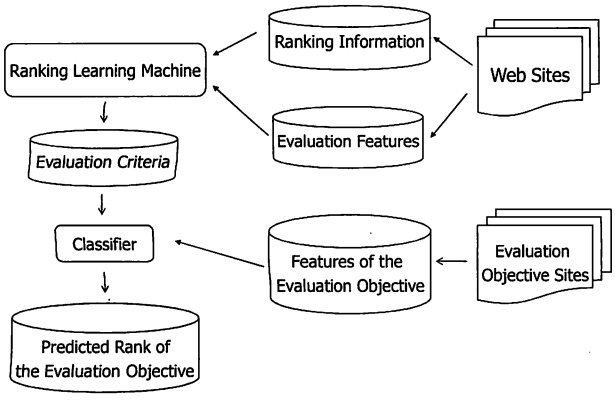

Fig. 4. System construction.

4. 実 験

本章では，〈3・1〉節で説明した3つのランキング学習アルゴリズムの比較実験を行い，本研究に適した学習手法を明らかにした。また，提案した評価属性を用いて抽出された評価基準の妥当性を検証するため，抽出精度の評価実験を行った。さらに，評価属性と分類結果の関係を調べるために属性の重要度を調べる実験を行った。そして，本研究の有効性を示すために，既存手法との評価精度の比較実験を行った。

評価データとして，4つの登録型ランキングサイトから集めてきた7カテゴリの計735サイトのランキング情報を用いた。その内訳は，SiteRankから384サイト，WEB RANKING[††]から128サイト，「人気サイトランキング」[†††]から185サイト，MATOMEX[††††]から38サイトである。ランク付けの方法としては，それぞれのカテゴリにおいて，Webサイトのランキングリストを5等分し，上位の方から1（優れている）～5（期待はずれ）の5段階に設定した。評価属性は〈3・2〉節で述べた31個の属性を用い，SVMツールはlibSVM2.84を使った。

〈4・1〉 評価尺度 情報推薦の分野で広く使われている統計的評価基準であるMAE (Mean Absolute Error) を用いて抽出精度を評価した。MAEの計算式は(6)式のようになる。

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad \cdots\cdots(9)$$

pは実際の評価値で，qはシステムの予測値，Nは評価されたアイテムの数である。MAEが低いほど，抽出の精度が高いことになる。

さらに，各手法の優劣をより詳細に見るため，MAEに加えて，予測値を実際の値に一致させるのに必要な最小隣接入れ替え数[22]を採り入れた。最小隣接入れ替え数とは，ランキング順位表で隣接した2つのWebサイトの順序を入れ替える最小の操作回数である。この評価尺度は，ランクのズレよりもさらに詳細な精度の評価ができる。

〈4・2〉 データの前処理とカーネル選択 評価属性の値は0から数百万とばらつきが大きい。そこで，前処理として，各属性に対して平均0分散1となるようにスケーリングを行った。libSVMが提供している学習カーネルはlinear，polynomial，radial basis function，sigmoidの4タイプである。これらのカーネルを使って予備実験を行い，それぞれの精度を調べた。その結果，精度の一番高かったRBFカーネルを実験で使用することとした。RBFカーネルには，cとγの二つのパラメータがある。パラメータの選択に当たって，前処理を行ったデータをlibSVMのグリッドサーチに

†† http://www.webranking.net/
††† http://ninkirank.misty.ne.jp/
†††† http://www.matomex.com/

Table 2. Experimental results of algorithm comparison.

	MAE	Minimum number of adjacent transpositions
Ranking SVM	0.78	132.7
Multi-class SVM	0.93	217.4
Prank	1.12	264.5

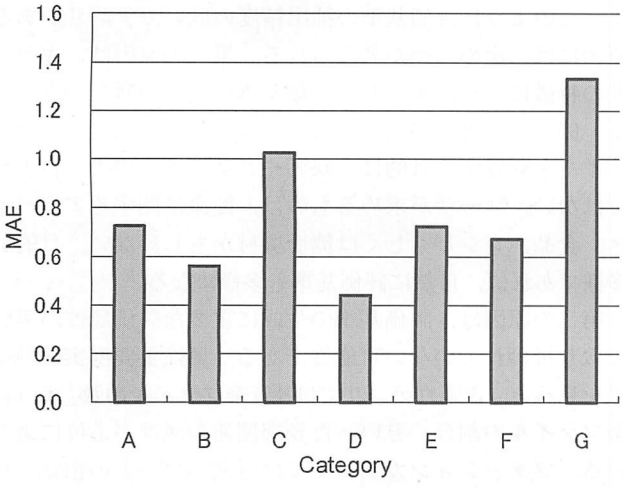

Fig. 5. Results of evaluation experiment. (The y-axis represent the values of MAE, and the x-axis correspond to the categories A: computer science, B: music, C: health and diet, D: movies, E: sports, F: games, G: fashion.)

かけ，カテゴリごとに最適なパラメータを決めた。

〈4·3〉 **学習アルゴリズムの比較実験** Ranking SVM, Prank とマルチクラス SVM の 3 つのランキング学習アルゴリズムに対して比較実験を行った。それぞれのカテゴリにおいて，10-fold クロスバリデーションを行い，MAE と最小隣接入れ替え数の平均値を表 2 に示す。表 2 から分かるように，Ranking SVM は他の 2 手法に較べて，MAE，最小隣接入れ替え数ともに優れたパフォーマンスを示している。一般的には，マルチクラス SVM は，ランキングのようなクラス間の順序関係を直接扱うことは難しい。Prank は順序関係のあるランキング学習は可能なものの，線形分離可能なデータを扱うことが前提となっているため，線形分離不可能なデータでは誤差が大きくなる可能性が高い。この実験結果から，Ranking SVM は，他の類似の学習アルゴリズムよりも本研究に適した手法であると結論できる。また，本研究で扱っているランキングの学習問題は，線形分離不可能であり，単純に解くことが困難な問題であることも分かる。

〈4·4〉 **抽出精度の評価実験** 本研究で提案した評価属性を用いて抽出された評価基準の妥当性を検証するため，抽出精度の評価実験を行った。学習アルゴリズムは〈4·3〉節で最もよい性能を示した Ranking SVM を用いる。訓練データの偏りに依存しないパフォーマンスを評価するために，カテゴリごとに 10-fold クロスバリデーションを行った。その結果を図 5 に示す。MAE の平均値は 0.78 で，標準偏差は 0.3 となった。つまり，ほとんどのカテゴリにおいて，システムの予測値はユーザの実際の評価値と一致しているか，1 ランク程度のズレであることがわかる。本研究と同じく 5 段階の評価尺度を用い，MAE を評価基準とした先行研究[23)~(25)]の実験結果を参考にすると，1 ランク以下のズレに抑えることができれば，有効性があると考えられる。したがって，本研究で提案した評価属性を用いて抽出された評価基準が妥当であると言える。

〈4·5〉 **属性の寄与度実験** ユーザや Web サイト作成

Table 3. Top 5 important features in each category.

	computer science
1	Indexable text information
2	Percentage of updated text information
3	Number of links(site)
4	Specification of author information
5	Top page's PageRank
	music
1	Number of links(top page)
2	Number of audio files
3	Number of links(site)
4	Top page's PageRank
5	Percentage of audio files
	health and diet
1	Indexable text information
2	Number of image files
3	PR for search robots
4	Specification of image size
5	Number of links(site)
	movies
1	Number of links(site)
2	Number of image files
3	Number of links(top page)
4	Percentage of image files
5	Specification of author information
	sports
1	Number of image files
2	Number of links(top page)
3	Indexable text information
4	Specification of background image
5	Top page's PageRank
	games
1	Number of links(site)
2	Number of image files
3	Specification of background image
4	Number of links(top page)
5	Indexable text information
	fashion
1	Indexable text information
2	Percentage of updated text information
3	Number of image files
4	Specification of cache expiration date
5	Specification of font code

者にとって，どの評価属性がどのように評価に影響を与えているかを知ることが重要である。本研究で扱うデータは線形分離不可能と考えられるため，Ranking SVM はカーネルトリックを用いている。しかし，カーネルトリックを適用した段階で，学習された判別関数から直接的に各分類属性の寄与率を調べることができない。そこで，属性選択（feature selection）で一般に使われる，属性の部分集合の評価を機械学習手法自身により行うラッパー法[26]を用いて属性の寄与度を調べた。代表的な探索アルゴリズムとしては，空集合から始めて一つずつ属性を追加していく前向き探索，全属性から始めて一つずつ集合から属性を取り除いていく後向き探索がある。この実験では後向き探索を用い，部分集合の評価については，MAE を基準にする。

それぞれのカテゴリにおいて，ラッパー法による属性選択により得られた重要度がもっとも高い上位 5 つの評価属性を表 3 にまとめた。表 3 から，カテゴリごとに各属性の重要度が異なること，多くのカテゴリにおいて，HTML テキストの量やリンクの数が評価に大きな影響を与えることが分かる。

〈4·6〉 既存の重要度推定手法との比較実験 本研究で提案した手法は，Web サイトの重要度をその評価関数の学習により適応的に推定している。その有効性を示すために，既存の重要度推定手法との比較実験を行った。比較対象として，一般的用いられる代表的な重要度推定手法である PageRank を選んだ。それぞれのカテゴリにおいて，PageRank で分類した場合と提案手法で分類した場合の分類精度を実験を通して示し，MAE により比較を行った。ここで PageRank による分類とは，Web サイトを PageRank の高い順に並べ，提案手法と同じようにランキングリストを 5 等分し，上位の方から 1（優れている）～5（期待はずれ）のランク付けを行うことである。実験結果を表 4 にまとめた。この実験結果から，提案手法はすべてのカテゴリにおいて，PageRank より分類精度が高いことが分かり，既存の固定された重要度推定手法よりも優れていることが確認された。

Table 4. Comparison of MAE with PageRank.

Category	Proposal	PageRank
Computer science	0.73	1.57
Music	0.56	2.12
Health and diet	1.03	1.86
Movies	0.44	1.73
Sports	0.72	2.23
Games	0.67	2.57
Fashion	1.33	1.48
Average MAE	0.78	1.94

5. 考　察

〈5·1〉 カテゴリの影響 図 3 の評価実験の結果を見ると，カテゴリごとの評価精度の差が小さくないことが分かる。結果が良かった B と D は音楽と映画カテゴリで，C は健康·ダイエット，最も悪かった G はファッションである。このように評価基準の抽出精度の低いカテゴリがある原因には，次の二つが考えられる。第一の原因は，カテゴリの特徴により，統一した評価基準の設定が難しい点である。例えば，健康·ダイエットカテゴリにおいて，ユーザがサイトを閲覧する目的は，映画や音楽サイトのように単純ではない。ユーザが求めるものは，健康に関するアドバイス，食品，レシピもしくは健康機材かもしれない。目的が多様であれば，自然に評価基準も多様になる。

第二の原因は，評価基準の生成に重要な評価属性が実験で取り扱われていない可能性である。属性重要度実験の結果を見ると，音楽カテゴリでは「音楽ファイルの数」や「音楽ファイルの割合」といった音楽関連の属性が上位に来ている。ファッションカテゴリでは「文字コードの指定」や「キャッシュ有効期限の指定」のようなファッションとあまり関係なさそうな属性が分類結果に大きな影響を与えている。この結果から，ファッションカテゴリにおいて，評価基準の生成に重要な評価属性が今回の実験では取り扱われていないと考えられる。現段階の Web テクノロジーでは取得不可能な評価属性，例えば「センスのいい美しい画像」が重要なのであれば，ファッションカテゴリの抽出精度を上げることは難しい。

〈5·2〉 訓練データの量と質 今回の実験では，7 カテゴリの 735 サイトを対象にしたが，必要なデータを揃えるために複数の登録型ランキングサイトからデータを収集している。しかし，登録型ランキングサイトごとにユーザ層や登録サイトの質が偏っている可能性があるため，抽出精度を低下させる原因になっているかも知れない。今後登録型ランキングサイトの規模がもっと大きくなり，一つのサイトから必要十分なデータを入手できるようになれば，抽出精度の向上が期待できる。また，登録型ランキングサイトではカテゴリがより細分化される傾向があり，個々のカテゴリのトピックがより明確になれば，評価基準もより明確になり，抽出精度も良くなるだろう。

〈5·3〉 評価属性とその寄与度 表 3 でリストアップした属性は，評価基準の抽出精度に大きな影響を与えている。しかし，精度にほとんど影響しない属性や，特定のカテゴリにおいて抽出精度を下げている属性も存在する。このような属性を取り除いたほうが，今回の実験においては有利になるが，評価基準は時間とともに変動するため，現段階では有効ではなかった属性が，いずれ有効な属性になる可能性も否定できない。したがって我々は，今後評価属

性を増やすことはあっても減らすことはない。〈3・2〉節で述べたように，Webロボットの改善やWebテクノロジーの発展に伴って，利用可能な評価属性も増えてくる。今回抽出精度が悪かったカテゴリでも，適切な属性が扱えるようになれば，精度の向上につなげることが期待できる。

6. むすび

Webのさらなる発展のためには，それらの量の増大に加えて質の向上が必要不可欠である。そのためには，Webサイトを客観的に評価し，ユーザに提示することが重要である。Webサイト評価は基本的には手動で行われるが，それでは処理できるWebサイトの規模に明らかに限界があり，コンピュータシステムによる自動的なWebサイトの評価が強く望まれる。

本研究では，既存の評価データからWebサイトの評価基準を自動抽出し，未評価のサイトに対して，抽出された評価基準に基づく自動評価を実現した。具体的には，ランキングサイトからWebサイトとそのランクを自動的に収集し，そのWebサイトを評価属性で記述したものを訓練データとする。そして，その訓練データからランキングの評価関数を学習することで，Webサイトの評価基準を抽出した。我々は，3つのランキング学習アルゴリズムに対して比較実験を行い，Ranking SVMがもっとも適した学習アルゴリズムであることを明らかにした。また，抽出精度の評価実験を行い，提案した評価属性を用いて抽出された評価基準が妥当であることを確認した。さらに，属性選択によりそれらの属性の寄与度を調べ，カテゴリ毎に異なる属性が大きく寄与することが分かった。そして，既存の重要度推定手法との比較実験を行い，本研究の有効性を確認した。

（平成21年5月11日受付，平成21年9月16日再受付）

文　献

(1) M. Richardson, A. Prakash, and E. Brill : "Beyond PageRank: machine learning for static ranking", Proc. 15th International Conf. on World Wide Web, pp.707-715 (2006-5)

(2) S. Ssemugabi and R. de Villiers : "A comparative study of two usability evaluation methods using a web-based e-learning application", Proc. 2007 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries, pp.132-142 (2007-10)

(3) A. Aizpurua, M. Arrue, M. Vigo, and J. Abascal : "Transition of accessibility evaluation tools to new standards", Proc. 2009 International Cross-Disciplinary Conf. on Web Accessibililty, pp.36-44 (2009-4)

(4) E. Velleman, C. Strobbe, J. Koch, C. A. Velasco, and M. A. Snaprud : "Unified Web Evaluation Methodology Using WCAG", Proc. 4th International Conf. on Universal Access in Human-Computer Interaction, Vol.4556, pp.177-184 (2007-8)

(5) L. Falk, A. Prakash, and K. Borders : "Analyzing Websites for User-Visible Security Design Flaws", Proc. 4th Symposium on Usable Privacy and Security, pp.117-126 (2008-7)

(6) O. Chapelle and Y. Zhang : "A dynamic bayesian network click model for web search rank", Proc. 18th International Conf. on World Wide Web, pp.1-10 (2009-4)

(7) S. Rajaram, A. Garg, X. S. Zhou, and T. S. Huang : "Classification approach towards ranking and sorting problems", Proc. of 14th European Conf. on Machine Learning, pp.301-312 (2003-9)

(8) K. Crammer and Y. Singer : "Pranking with ranking", Advances in Neural Information Processing Systems 14, Vol.1, pp.641-647, MIT Press (2002)

(9) E. L. Allwein, R. E. Schapire, and Y. Singer : "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers", Journal of Machine Learning Research, Vol.1, pp.113-141 (2000)

(10) J. E. Alexander and M. A. Tate : "Web WISDOM: How to evaluate and create information quality on the Web", Lawrence Erlbaum, Hillsdale, NJ (1999)

(11) M. P. Arnone and R. V. Small : "WWW motivation mining finding treasures for teaching evaluation skills grade 1-6", Linworth Publishing, Worthington, OH (1999)

(12) K. Sumi and A. Yotsuya : "A Scale to Assess Website as Information Resource in Seeking Needed Information: Construction, Reliability and Validity", IPSJ Japan, Vol.45, No.3, pp.1032-1040 (2004) (in Japanese)
鷲見克典・四谷あさみ：「調べる目的で利用する情報源としてのWebサイトに対する評定尺度の作成と信頼性および妥当性の検討」，情処学論，**45**, 3, pp.1032-1040 (2004)

(13) M. Y. Ivory : "An Empirical Approach to Automated Web Site Evaluation", Journal of Digital Information Management, Vol.1, No.2, pp.75-102 (2003)

(14) G. Velayathan and S. Yamada : "Behavior-based Web page evaluation", Journal of Web Engineering, pp.222-243 (2007)

(15) L. Olsina, D. Godoy, G. J. Lafuente, and G. Rossi : "Specifying quality characteristics and attributes for Websites", First ICSE Workshop on Web Engineering, pp.266-278 (1999)

(16) E. K. R. E. Huizingh : "The Content and Design of Web Sites: An Empirical Study", Information and Management, Vol.37, No.3, pp.123-134 (2000)

(17) M. B. Mateos, A. C. Mera, F. J. M. Gonzalez, and O. R. G. Lopez : "A new Web assessment index: Spanish universities analysis", Internet Research: Electronic Networking Applications and Policy, Vol.11, No.3, pp.226-234 (2001)

(18) W. P. Palmer : "Web Site Usability, Design, and Performance Metrics", Information Systems Research, Vol.13, No.2, pp.151-167 (2002)

(19) R. Sinha, M. Hearst, and M. Y. Ivory : "Content or Graphics?: An Empirical Analysis of Criteria for Award-Winning Websites", Proc. 7th Conf. on Human Factors and the Web (2001-6)

(20) M. Y. Ivory and M. A. Hearst : "Statistical Profiles of Highly-Rated Web Sites", Proc. SIGCHI Conf. on Human Factors in Computing Systems, pp.367-374 (2002-4)

(21) X. Qi and B. D. Davison : "Web page classification: Features and algorithms", ACM Computing Surveys, Vol.41, No.2, pp.1-31 (2009-2)

(22) G. Lebanon and J. Lafferty : "Cranking: Combining rankings using conditional probability models on permutations", Proc. 19th International Conf. on Machine Learning, pp.363-370 (2002-7)

(23) H. Takashima, H. Yamagishi, and S. Hirasawa : "An Improved Method of Collaborative Filtering with Predicting Unobserved Values", FIT Japan, A-008 (2005-9) (in Japanese)
高島秀佳・山岸英貴・平澤茂一：「欠損値推定による協調フィルタリング手法」，情報科学技術フォーラム論文集，A-008 (2005-9)

(24) B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl : "Item-based collaborative filtering recommendation algorithms", Proc. 10th International Conf. on World Wide Web, pp.285-295 (2001-5)

(25) P. Li and S. Yamada : "A Movie Recommender System Based on Inductive Learning", Proc. of IEEE Conf. on Cybernetics and Intelligent Systems, Vol.1, pp.318-323 (2004-12)

(26) R. Kohavi and G. H. John : "Wrappers for Feature Subset Selection", Artificial Intelligence, Vol.97, No.1-2, pp.273-324 (1997)

李　鵬　（非会員）　1978年11月16日生。2002年3月千葉工業大学情報学部卒業。*2004*年3月東京工業大学総合理工学研究科修士課程修了。同年4月同大学総合理工学研究科知能システム科学専攻博士後期課程入学，現在に至る。人工知能，特にWebインテリジェンスやデータマイニングに興味を持つ。

山 田 誠 二

（非会員） 1960 年 10 月 11 日生。1984 年大阪大学基礎工学部卒業。 1989 年同大学院基礎工学研究科博士課程修了。工学博士。1989 年大阪大学基礎 工学部助手。1991 年大阪大学産業科学研究所講師。1996 年東京工業大学大学院 総合理工学研究科助教授。2002 年国立情報学研究所教授，現在にいたる。知的 Web インタラクション，HAI の研究に従事。情報処理学会，日本ロボット学会，AAAI, IEEE, ACM，各会員。

新 田 克 己

（非会員） 1952 年 11 月 2 日生。1975 年東京工業大学工学部電子工学科卒業。1980 年同大学院理工学研究科博士課程修了。工学博士。1980～95 年電子技術総合研究所に勤務。その間，1989～94 年の間に（財）新世代コンピュータ技術開発機構に出向。1996 年東京工業大学大学院総合理工学研究科教授，現在に至る。主な研究内容は，論理プログラミング，エキスパートシステム，マルチエージェント，ヒューマンインタフェースなど。特に，法律分野への応用に興味をもつ。