DEIM Forum 2018 D2-2

HTML 構造上の特徴を利用した 学術用語解説ウェブページの分かり易さの自動評定

春日　孝秀† 塩川　隼人†† 韓　炳材†† 宇津呂武仁††† 河田　容英††††

† 筑波大学 理工学群 工学システム学類　〒 305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学 大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
††† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
†††† (株) ログワークス　〒 151-0053 東京都渋谷区代々木 1-3-15 天翔代々木ビル 6F

あらまし　本論文では，ウェブ上で学術用語を解説するページ群を対象として，それらのページ群における用語解説の分かり易さを自動評定する手法を提案する．ウェブページの分かり易さ評定を，分かり易い用語解説ウェブページが満たすべき 6 個の個別因子へ分解する．理工系学術分野を対象とし，学術用語を検索クエリとする検索上位ウェブページを収集し，各ページに対し人手により個別因子の評定及び全体評定を行い，参照用データの作成を行う．本論文では，特に，ウェブページの HTML 構造上の特徴量を利用した分類器学習により，用語解説ウェブページの分かり易さが評定可能であることを示す．

キーワード　学術用語解説ウェブページ，学習教材読解支援システム，ウェブ学習，検索エンジン，HTML 解析，ウェブページ収集，SVM

Measuring Beginner Friendliness of Web Pages explaining Academic Concepts based on HTML Structures

Takahide KASUGA†, Hayato SHIOKAWA††, Bingcai HAN††, Takehito UTSURO†††, and Yasuhide KAWADA††††

† College Eng. Sys., School Sci. and Eng., University of Tsukuba,　Tsukuba, 305-8573, Japan
†† Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba 305-8573 Japan
††† Faculty of Engineering, Information and Systems, University of Tsukuba,　Tsukuba 305-8573 Japan
†††† Logworks Co., Ltd.　Tokyo 151-0053, Japan

1. はじめに

インターネットが普及した現在，ウェブ上には多くの学術関連のコンテンツが存在する．このコンテンツを使って学術用語を学ぶことは，学校で使用される教科書・参考書による学習とは別に，学習手段の一つとしてたびたび用いられている．インターネットで学術用語を学ぶ際，ウェブ検索に頼れば関連ページは容易に見つかる．しかし，「分かり易い用語解説」を見つけるには，検索上位ページを一つずつ見比べ読み進める非効率な作業が必要となる．これは，ウェブ検索では，分かり易さをベースにした解説ページ群の体系化ができないことが原因である．そこで，本研究では，検索上位ページの中で分かり易さを充足する必要十分な数の用語解説ページを見つけ体系化することを目的とする．さらに，この研究においては，最終的に，用語解説を提示してウェブ学習を促進する学習教材読解支援システムを実現する．

以上の目的をふまえて，本論文では，図 1 に示すように，6 個の個別因子を手がかりとして，分かり易さの全体評定を行う手法を構築する．特に，本論文では，用語解説ウェブページの HTML 構造を素性とする分類器学習により，全体評定の自動判定を行う．具体的には，理工系学術用語を検索クエリとして収集した用語解説ウェブページに対して人手による評定を行った．そして，そのうちの約 360 ページを参照用事例として評価実験を行い，提案手法の有効性を評価する．

2. 用語解説ウェブページの「分かり易さ」評定の因子

本節では，用語解説ウェブページの個別因子および全体評定

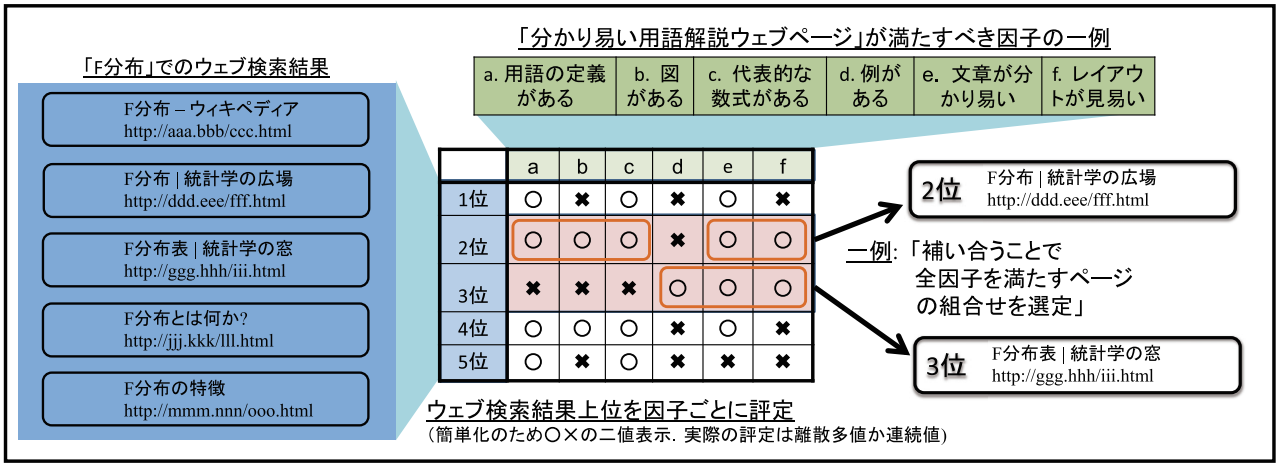


図 1　用語解説ウェブページの「分かり易さ」評定の因子

の評定基準について述べる．これらの評定においては，本論文の著者が二値での評定を行った．

2.1　個別因子

(a)　用語の定義がある

必要十分な用語の定義を述べている．例えば，統計分野の用語を検索クエリとしたとき，エクセル操作の解説ページがあり，学術用語に関する説明がされず，応用的な内容のみであれば定義無しとする．

(b)　図がある

図 2 の例に示すように，用語解説や例示の為に使用されている図を一つ以上含む．文字のみの表などは対象としない．

(c)　数式がある

図 3 の例に示すように，一般的に数式と呼べるものを含む．アルファベット一文字等は含めない．

(d)　例がある

図 4 の例に示すように，例題や,「例えば，～のように」などの例示を含む．例示については，用語を説明するための短すぎない例示に絞る．一つでも例題，例示と判断できるものがあれば良いものとした．

(e)　レイアウトが見易い

ページのレイアウトの見易さ．学術用語解説ウェブページの見易さに特化する，というよりは，一般的なウェブページの見易さに準ずる．例えば，シンプルさ，改行や章題による構成の工夫，色遣い等が挙げられる．

(f)　文章が分かり易い

用語解説ページの文章の分かり易さ．そのページで扱う用語と，対象とするページ閲覧者層を考慮し，総合的な評価で行う．具体的には，例えば，専門用語と平易な用語のバランス，情報量，文章量等を基準に判断を行った．

2.2　全体評定

6 個の個別因子を元に，ウェブページ全体の評価付けを行う．全体評定の評価は，個別因子の評定がどの程度満たされているかを基準として判断しているが，個別因子全てが充足されている必要はなく，ある程度充足されていれば，全体評定も充足すると評定している．例えば,「図あり」や「式あり」などの因子は必ずしも必要ではなく，これらの因子が欠けていたとしても全体評定が充足されないわけではない．6 個の因子の中で特に重要なのは「レイアウトが見易い」と「文章が分かり易い」のため，この両方が欠けている場合，あるいはどちらか片方が欠けていて，かつ，その他の個別因子が複数欠けている場合は，全体評定を「充足しない」と評定している．

3.　参照用学術用語解説ウェブページ集合の作成

3.1　対象学術分野および用語

事前調査を行い,「レイアウトの見易さ」,「文章の分かり易さ」,「全体評定」の評定基準が似ている統計・線形代数・物理分野を本研究での対象分野として選んだ．次に，各分野の用語の内，主に高校 3 年生から大学の学校教育レベルのものの中から，表 1 に示すように，統計のみ 30 語，線形代数・物理は 15 語を対象学術用語として選んだ．統計のみ 30 語で他の 2 分野より多いが，理由としては統計分野が他の学術分野よりも経済的な側面が強く，実際に経済分野で多く利用され，検索数が多い用語が他の分野に比べて多いためである．

3.2　作成手順および結果

前節で収集した学術用語に対して用語解説ウェブページを収集し，2. 節で述べた個別因子および全体評定を人手で判定した事例を蓄積する．

まず，ウェブページの収集には検索エンジンを用い，各用語を検索クエリとした場合の検索結果上位 10 件のウェブページを集めた．その際，HTML ファイル収集プログラムではアクセスできないページは参照用ページ集合に加えない．また，サイト数としては少数であるが，そのサイトのページが多数上位 10 位以内に含まれるサイトのページ (質問回答サイトのページ，辞書サイトのページ，Wikipedia のページ等) も参照用ページ集合に加えない．以上の手順により，収集対象ページ集合の部分

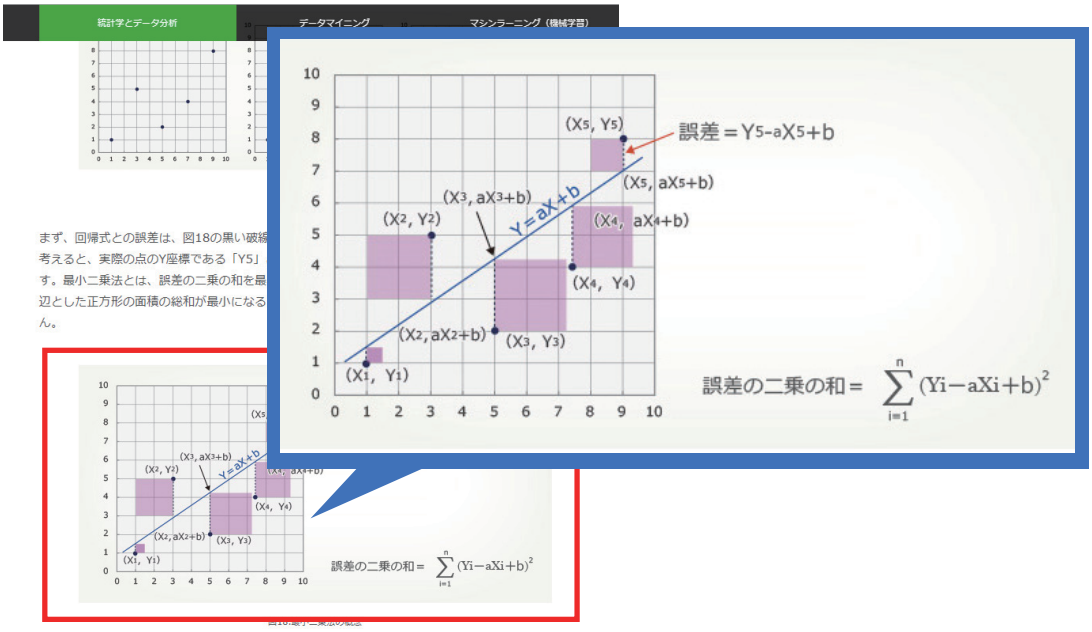


図 2　用語解説ウェブページにおける図の例 (https://www.albert2005.co.jp/knowledge/statistics_analysis/multivariate_analysis/single_regression より抜粋)

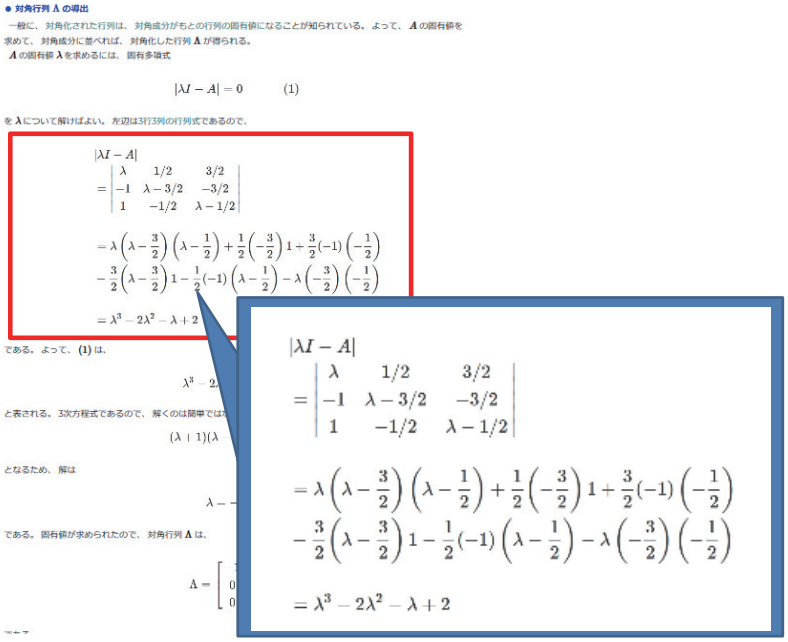


図 3　用語解説ウェブページにおける式の例 (http://physmath.main.jp/src/matrix-diagonalization-example-3-3.html より抜粋)

集合として，参照用ページ集合を作成した．なお，参照用ページ集合中には，数%程度用語解説以外のページが含まれる．以上の手順によって収集された参照用ページの数を表 2 に示す．

次に，収集対象の全ページに対して個別因子の評定および全体評定を人手で行った．この結果について，収集対象の全ページ中の統計，および，参照用ページ中の統計をそれぞれ表 3 に示す．

4. HTML 構造上の特徴を素性とする分類器学習による個別因子及び全体評定の判定

本節では，用語解説ウェブページに対する 6 個の個別因子および全体評定を対象として分類器学習手法を適用し，評価を行う．

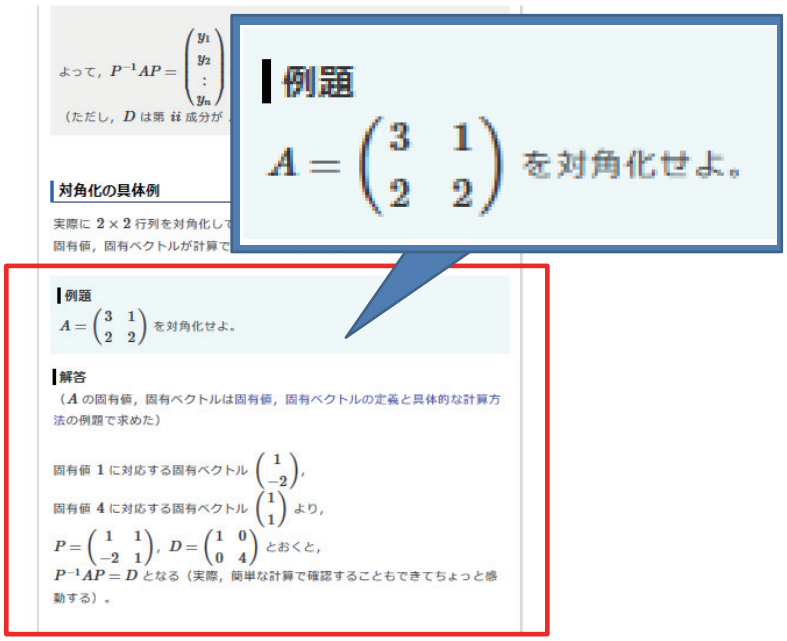


図 4　用語解説ウェブページにおける例示の例 (https://mathtrain.jp/diagonalization より抜粋)

表 1　対象とする学術用語リスト

分野	対象とする学術用語数	対象とする学術用語
統計	30	正規分布，F 分布，信頼区間，標準偏差，主成分分析，回帰分析，自己回帰モデル，ポアソン分布，マルコフ連鎖モンテカルロ法，確率密度関数，コーシー分布，分散，確率，事後分布，事前分布，相関係数，共分散，自己回帰，中央値，ガンマ分布，ノンパラメトリック手法，パレート分析，十分統計量，ホワイトノイズ，独立成分分析，多変量解析，帰無仮説，ラプラス分布，モーメント，k 平均法
線形代数	15	階数，共役勾配，行列式，クラメルの公式，クロネッカーのデルタ，三角行列，正規直交基底，対角化，直交行列，特性多項式，二次形式，ノルム，メネラウスの定理，ヤコビ行列，内積
物理	15	電気力線，張力，慣性の法則，遠心力，電波，電流，万有引力，交流，音波，ホイートストンブリッジ，反発係数，相互誘導，正電荷，速度，変圧器
合計	60	—

表 2　対象クエリ数・収集対象ページ総数・参照用ページ数

分野	クエリ数	収集対象ページ総数	参照用ページ数
統計	30	300	181
線形代数	15	150	78
物理	15	150	96
合計	60	600	355

4.1　素　　性

素性として，表 4 に示すように，HTML タグ部分，および，テキスト部分を対象とする二種類の素性を用いる．

4.1.1　HTML 素性

HTML 素性としては，body 要素の高さ，各 HTML タグのページ内延べ使用頻度，および，各 HTML タグ属性のページ内延べ使用頻度を値として用いる．各 HTML タグおよび HTML タグ属性ごとに素性としての有効性を評価し，表 5 において「素性として利用」として列挙したもののみを利用する．

4.1.2　テキスト素性

HTML タグ部分以外のメニューバー部分，ページ最下部の管理者情報等も含む全テキスト部分を対象として，表 4 に示す各種パターンを素性とする (注1)．

(注1)：テキスト部分の抽出においては，Python ライブラリの readability を用いた．

表 3　参照用学術用語解説ウェブページ集合の個別因子及び全体評定の充足率 (%) (参照用ページ中 / 収集対象ページ中)

分野	定義あり	図あり	式あり	例あり	見易い	分かり易い	全体評定
統計	88 / 91	67 / 65	71 / 77	74 / 63	45 / 49	46 / 48	30 / 32
線形代数	94 / 92	33 / 27	97 / 91	68 / 53	39 / 26	51 / 38	44 / 27
物理	58 / 77	49 / 42	41 / 43	32 / 32	31 / 27	38 / 41	20 / 17
合計	81 / 88	55 / 50	69 / 72	61 / 50	40 / 38	45 / 44	30 / 27

表 4　素 性 一 覧

種類	素性名	説明	値
HTML 素性	HTML 高さ	body 要素の高さ (px)	多値
	HTML タグ	当該 HTML タグのページ内延べ使用回数	多値
	HTML タグ属性	当該 HTML タグ属性のページ内延べ使用回数	多値
テキスト素性	文字種	各ページ本文内のひらがなの数から漢字の数を引いた文字数の差	多値
	「！」文字数	各ページの HTML 内のエクスクラメーションマークの数	多値
	「プライバシーポリシー」有無	各ページの HTML 内に,「プライバシーポリシー」「privacy policy」のどちらかの文字列が 1 つでもある	2 値
	「ホーム」有無	各ページの HTML 内の,「ホーム」という文字列があるかどうか	2 値

表 5　HTML タグおよび HTML タグ属性の一覧

	素性として利用	素性として利用せず
HTML タグ	iframe, div, noscript, meta, nav, ul, li, article, h1, h2, h3	input, script, strong, form, src, table, section, h4, h5, button
HTML タグ属性	onload, sizes, data-share, required	alt, media, scrolling, data-layout, frameborder, method, data-ad-slot, data-remodal-action, role, itemtype, async

4.2 評 価 手 順

分類器としては，scikit-learn [2] パッケージにおける SVM(sklearn.SVM.SVC ツール) を用いる．評価においては，分野およびクエリを考慮せず無作為に混合し，8 分割交差検定により評価を行う．3 値以上の多値をとる素性については，素性のとる値の分布に応じて，重複のない多数の範囲をとる二値素性，および，重複のある多数の範囲をとる二値素性等，様々な二値素性のとり方を評価した．その結果，左端を固定し，右端が異なる 20〜40 個の範囲をとる二値素性を用いた場合に最も高い性能となったため，この素性設定を用いた．

表 3 に示すように，参照用データ中の「全体評定」における正例・負例比は約 3:7 であるので，評価においては以下の手順を 7 回繰り返し，その平均を求めて評価結果とした．

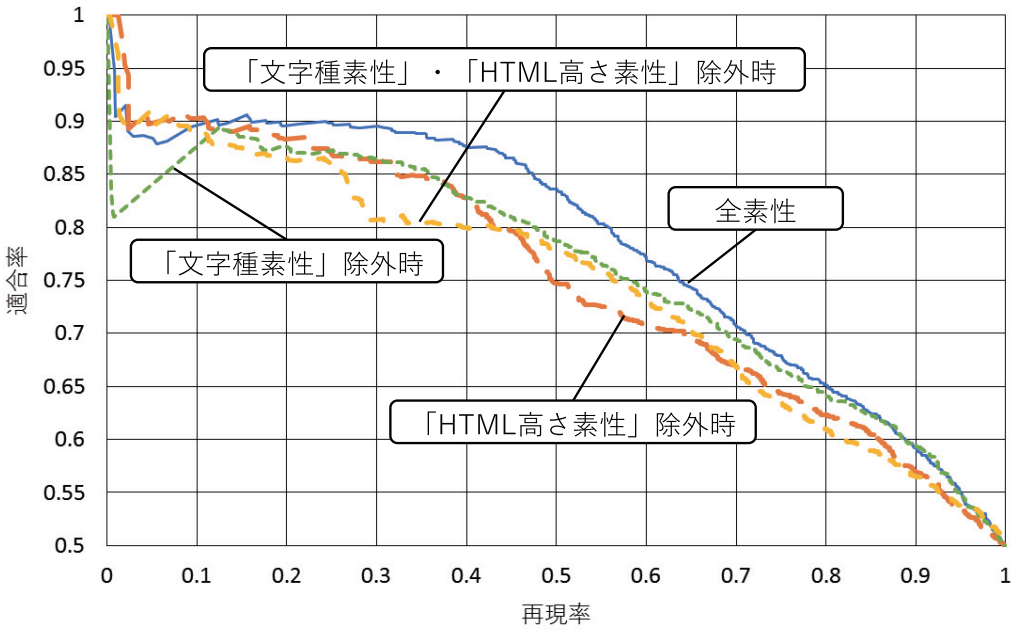
（1） 正例集合を P とする．負例を無作為に 3:3:1 に分割し，負例集合 N_1，N_2，N_{rest} とする．

（2） 正例・負例比が 1:1 となる正例・負例集合を，$P \cup N_1$，および，$P \cup N_2$ の二セット用意し，それぞれにおいて 8 分割交差検定を行い，その評価結果の平均を求める．

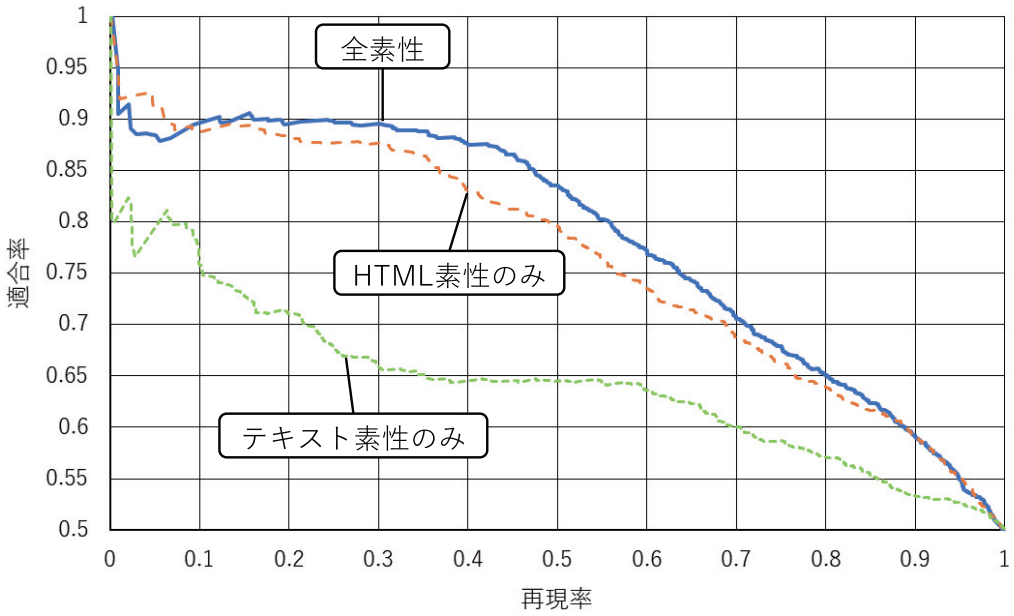
SVM のカーネル関数としては RBF カーネルを用い，最適化の目的関数を再現率・適合率の ROC 曲線の面積として，グリッドサーチによりコストパラメータ C(1 および 10) および RBF カーネルのパラメータ γ(0.01，0.001，および，0.0001) を最適化する．

4.3 評 価 結 果

評価においては，まず，表 4 に示す各種素性の各々を除外した場合の性能の低下度合いを測定することにより，各種素性の有用性を評価した．その結果，有用な素性として，HTML 高さ素性および文字種素性が得られた．そこで，これらの片方，もしくは，両方を除外した場合，および，全素性を用いた場合の計 4 本のプロットを比較した結果を図 5(a) に示す．この結果から分かるように，これらの二素性は，それぞれ片方を除外した場合でも一定の性能低下が観測されるが，二素性両方を除



(a) HTML 高さ素性・文字種素性除外時



(b) HTML 素性・テキスト素性の各 1 素性のみでの訓練・評価時

図 5　HTML 構造上の特徴を素性とする分類器学習による全体評定の判定: 評価結果

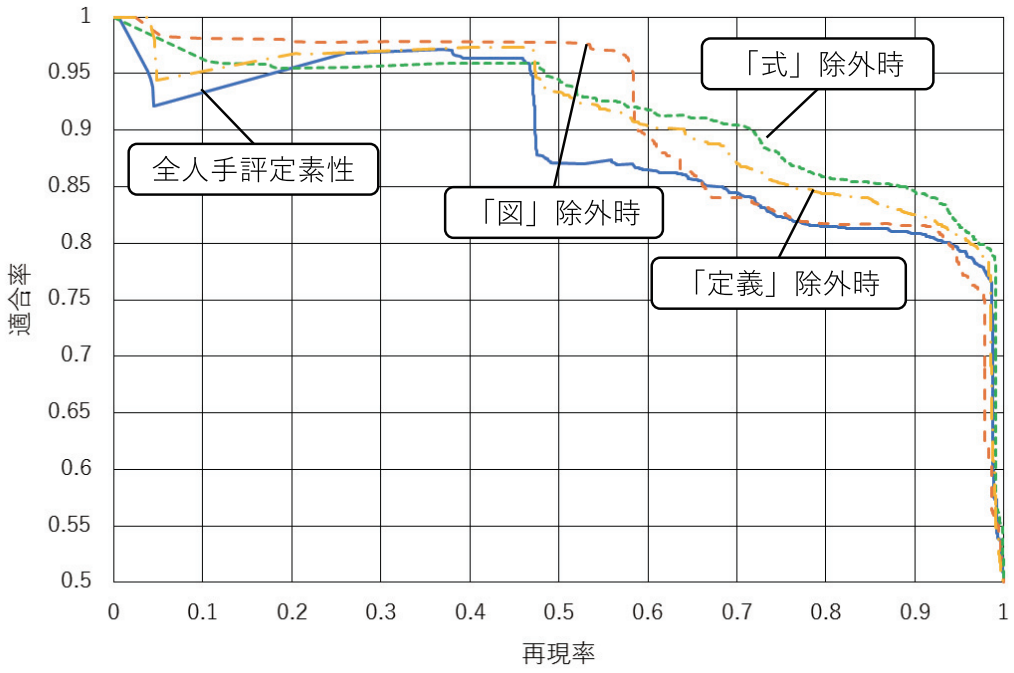
外した場合に更なる性能低下が観測されることが分かる．

次に，HTML 素性とテキスト素性のどちらの有用性が大きいのかを評価するために，各々片方の素性のみで訓練・評価を行った結果を図 5(b) に示す．この結果から分かるように，各素性単独の場合には，HTML 素性の性能の方が高いことが分かる．しかし，いずれの素性においても，全素性よりも性能が下がっていることから，この両素性は相互に補い合って性能向上に寄与していることが分かる．

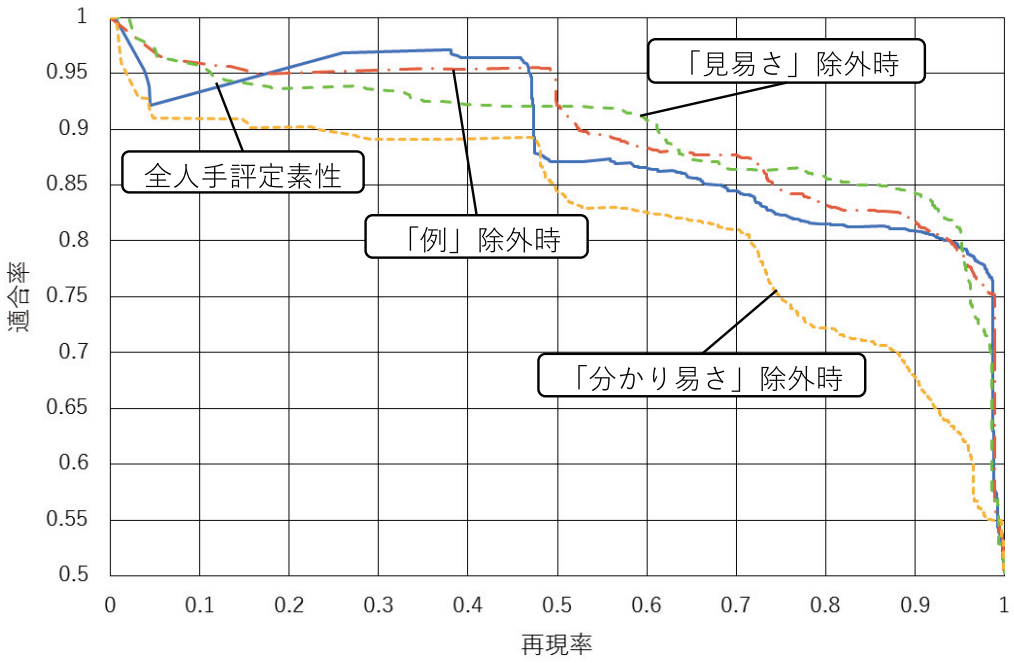
5. 人手による個別因子評定結果を素性とする分類器学習による全体評定の判定

本節では，用語解説ウェブページの個別因子の人手評定結果を素性とし，全体評定を判定対象のクラスとして分類器学習手法を適用する．分類器学習における評価手順は前節の手順に従う．いずれの素性も二値素性であるため，多値素性の二値素性への置き換えの手順は用いない．

図 6 に，人手評定を行った個別因子 6 素性をすべて用いた場合 (全人手評定素性)，および，6 素性のうちのそれぞれ 1 素性



(a) 定義，図，式除外時



(b) 例，見易さ，分かり易さ除外時

図 6　人手による個別因子評定結果を素性とする分類器学習による全体評定の判定: 評価結果

のみ除外した場合の，合計 7 本のプロットを示す.「文章の分かり易さ」を除外した場合 (「分かり易さ」除外時) において，最も大きく性能が低下しており，有用性が最も高い素性であることが分かる.

6. 関 連 研 究

学術用語解説ウェブページの分かり易さ自動評定タスクに関連して，コミュニティ型質問応答の分野においては，回答者による回答の良質さを自動評定する手法についての研究が行われている [1,3]．また，共同研究者の研究 [4] においては，本論文で対象とした個別因子のうち，特に,「レイアウトの見易さ」の因子を対象として，画像情報を特徴量とする深層学習による判定性能の評価を行っており，今後の課題として，この研究で提示された手法を組み合わせることによって，本研究の目的である用語解説ページの体系化を行うことが挙げられる.

7. おわりに

本論文では，6 個の個別因子を手がかりとして，分かり易さの全体評定を行う手法を構築した．特に，用語解説ウェブページの HTML 構造を素性とする分類器学習により，全体評定の自動判定を行った．具体的には，理工系学術用語を検索クエリとして収集した用語解説ウェブページに対して人手による評定を行った．そして，そのうちの約 360 ページを参照用事例として評価実験を行い，提案手法の有効性を評価した．

文　献

[1] 石川大介, 酒井哲也, 関洋平, 栗山和子, 神門典子. コミュニティ QA における良質回答の自動予測. 情報知識学会誌, Vol. 21, No. 3, pp. 362–382, 2011.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.

[3] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community QA answer selection. In *Proc. 4th WSDM*, pp. 187–196, 2011.

[4] 塩川隼人, 春日孝秀, 韓炳材, 宇津呂武仁, 河田容英. 深層学習を用いた学術用語解説ウェブページの見易さの自動評定. 第 10 回 DEIM フォーラム論文集, 2018.