The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

異種の素性を併用した 学術用語解説ウェブページの分かり易さの自動評定

Measuring Beginner Friendliness of Web Pages explaining Academic Concepts by Integrating Heterogeneous Features

塩川 隼人[*1] Hayato Shiokawa　春日 孝秀[*1] Takahide Kasuga　川口 輝太[*1] Kota Kawaguchi　韓 炳材[*1] Bingcai Han　宇津呂 武仁[*2] Takehito Utsuro　河田容英[*3] Yasuhide Kawada

[*1]筑波大学大学院システム情報工学研究科 Grad. Sch. Sys. & Inf. Eng, Univ. of Tsukuba　[*2]筑波大学システム情報系 Fclty. Eng, Inf. & Sys, Univ. of Tsukuba　[*4](株) ログワークス Logworks Co., Ltd.

Search engine is an important tool of modern academic study, but the results are lack of measurement of beginner friendliness. In order to improve the efficiency of using search engine for academic study, it is necessary to find a method of measuring the beginner friendliness of a Web page explaining academic concepts and to build an automatic measurement system. This paper studies how to integrate heterogeneous features including those generated from the image of the Web page by a variant of CNN as well as those extracted from the body text of the HTML file of the Web page. Integration is performed through the framework of the SVM classifier learning. Evaluation results show that heterogeneous features perform better than each individual type of features.

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，6 個の個別因子のうち「レイアウトの見易さ」についての因子を対象とし，深層学習を用いてこの因子を自動評定する方式 [塩川 18] によって得られる特徴量，および，用語解説ウェブページの HTML ソーステキスト中のテキスト情報の特徴量 [春日 18] を併用して，分類器として SVM を用いて全体評定の自動判定を行う手法 [春日 18] を適用し，異種の特徴量を効果的に併用可能であることを示す．特に，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を対象として行った評価結果をふまえて，提案手法の有効性を示す．

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．

連絡先: 塩川 隼人，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, 029-853-5427

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

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

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

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

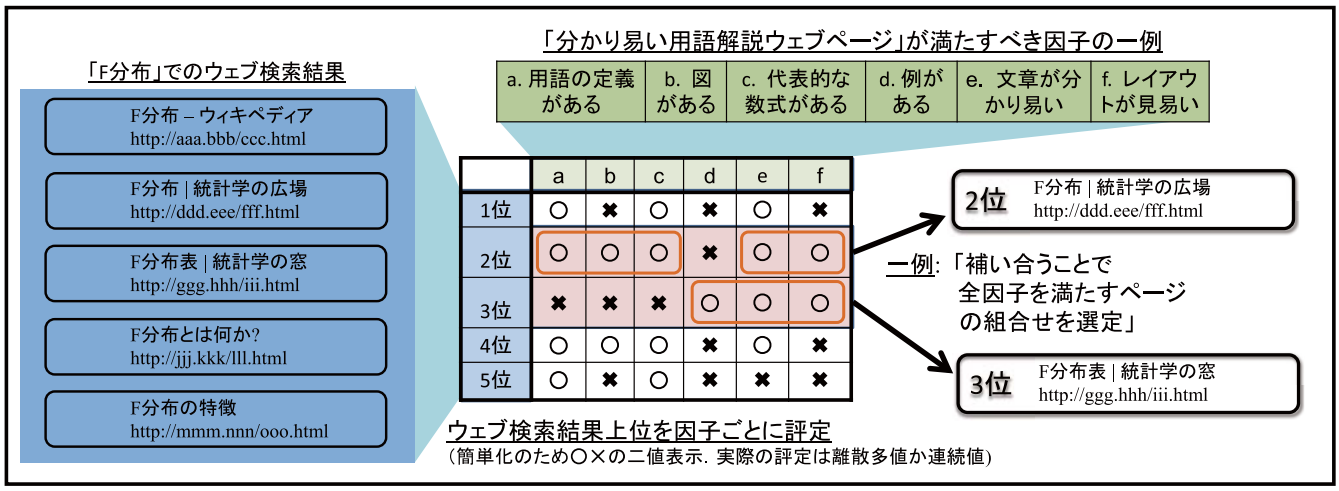

図 1: 用語解説ウェブページの「分かり易さ」評定の因子

表 2: 対象学術分野ごとの正例数及び負例数

	対象学術分野	正例数	負例数	合計
訓練事例	線形代数	36	54	89
	物理	26	101	127
	生物	17	47	64
	プログラミング	70	72	142
	IT	39	50	89
	小計	188	323	511
評価事例	統計	66	84	150
	化学	52	98	150
	合計	340	620	960

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理工系学術分野である統計・線形代数・物理・生物・化学・プログラミング・IT 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・物理・化学・プログラミング・IT は 15 語，線形代数のみ 14 語，生物のみ 7 語を対象学術用語として選んだ．

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．そして，2.2 節で述べた基準に基づき全体評定の判定を人手で行い，その判定結果を付与した事例を蓄積する．以上の手順によって，全体評定が「充足する」と判定された事例を正例,「充足しない」と判定された事例を負例として，各分野ごとに表 2 に示す数の正例・負例を収集した．全 7 分野のうち 5 分野を訓練用分野，残り 2 分野を評価用分野とした．

4. 深層学習による学術用語解説ウェブページの見易さの自動評定結果の付与

4.1 VGG16 [Simonyan15]

近年，深層学習は多様な分野のタスクにおいて，その有用性を示している．特に画像認識の分野では，畳み込みニューラルネットワーク (CNN) と ImageNet などの大規模なデータセットの登場により，様々なタスクにおいて驚異的な性能を発揮している．また，ImageNet のような大規模なデータセットを用いて学習を行った CNN のパラメータは，その汎用性の高さから高性能な特徴抽出器として異なるドメインのタスクにおいても活用できることが知られている．

そこで，本論文では，用語解説ウェブページの「レイアウトの見易さ」を評定するタスクにおいて，VGG16 モデル [Simonyan15] を基盤の特徴抽出器として用いる．VGG16 は ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 [Russakovsky14] の「画像分類タスク (Image classification)」で準優勝,「位置特定タスク (Single-object localization)」で優勝したモデルであり，このモデルは 13 層の畳み込み層，その途中に挟まれている 5 層の最大プーリング層，3 層の全結合層 (1,000 値分類用)，および，出力層から成り立っている．訓練済みの VGG16 としては，ImageNet2014 のデータセットによって 1,000 値分類用に訓練済みのモデルが一般に公開されており，この訓練済みモデルは他のタスクにも広く転用可能な

表 1: 学術分野ごとの対象学術用語一覧

分野	対象とする学術用語数	対象とする学術用語
線形代数	14	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 (計 14 語)
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 (計 15 語)
生物	7	DNA，ショウジョウバエ，原核生物，減数分裂，光合成，細胞，葉緑体 (計 7 語)
プログラミング	15	C 言語，Java，エスケープシーケンス，コマンドライン引数，スコープ，フィールド値，ポインタ，メソッド，繰り返し処理，構造体，算術演算子，条件分岐，配列変数，文字列，論理演算 (計 15 語)
IT	15	API，DBMS，HTML，IP アドレス，JDBC，RDB，SDK，SQL，Unicode，URL，スコープマネジメント，ステークホルダーマネジメント，タイムマネジメント，ナレッジマネジメント，リスクマネジメント（計 15 語）
統計	15	事後分布，確率，相関係数，信頼区間，k 平均法，回帰分析，十分統計量，確率密度関数，ガンマ分布，事前分布，主成分分析，コーシー分布，f 分布，自己回帰モデル，自己回帰 (計 15 語)
化学	15	イオン結合，エステル，カルボン酸，ケトン，化学反応式，化学平衡，共有結合，合成高分子，酸化還元，遷移元素，典型元素，天然高分子，燃料電池，物質の三態，芳香族有機化合物 (計 15 語)
合計	96	—

ことが知られている．本論文では，Python の深層学習ライブラリ Keras[*1] 用に公開されているモデル[*2] を利用して評価実験を行った．

4.2 見易さの自動評定結果の付与手順

本論文では，ImageNet データセットを用いて訓練済みの VGG16 モデルを基盤に用いて用語解説ウェブページの「レイアウトの見易さ」の判定の評価実験を行った．参照用学術用語解説ウェブページデータとして，3. 節で作成した用語解説ウェブページ事例を用い，各ウェブページのトップページを画像化したものを実際の事例として用いる．この訓練・評価事例集合の正例・負例数は，表 2 の通りである．

本論文において VGG16 を利用して「レイアウトの見易さ」を評定するモデルを訓練する際には，まず，VGG16 の 1,000 値分類用の 3 層の全結合層と出力層を取り払い，2 クラス分類用の全結合層 3 層と出力層を接続し，以下のモデルを用意する．

> 「fine-tuning-VGG16 ありのモデル」
>
> VGG16 の畳み込み層 10 層，および，最大プーリング層第 4 層目までのパラメータを VGG16 の訓練済みのパラメータに固定し，それ以降の畳み込み層 3 層，最大プーリング層 1 層，全結合層 3 層に対して，本論文の「レイアウト見易さ」判定タスク用の訓練事例を用いて，再度パラメータの訓練を行う．

対象学術分野のうち，統計および化学分野に属する学術用語をクエリとして収集した参照用学術用解説ウェブページ集合を評価事例とし，それ以外の事例を訓練事例とする．そして，以下の手順で，訓練・評価事例に対して深層学習による見易さの自動評定結果を付与する．

訓練事例 全 5 分野のうちの 4 分野で訓練したモデルを用いて，残りの一分野の事例に対して見易さの自動評定結果およびその信頼度を付与する．

評価事例 訓練事例の全 5 分野のうちの 4 分野で訓練したモデルを全 5 通り作成し，各評価事例に対して 5 つのモデルのうちの一つを無作為に適用し，見易さの自動評定結果およびその信頼度を付与する．

5. 深層学習による見易さ自動評定結果およびテキスト特徴量を素性とする分類器学習による全体評定の判定

5.1 素性

5.1.1 深層学習による見易さ自動評定結果の素性

4.2 節で付与された見易さの自動評定結果およびその信頼度の情報は，[0,1] の区間の数値情報として付与されており，信頼度=0.5 の場合を中間点として，値が 1 に近いほど，「自動評定=見易い」の信頼度が高く，値が 0 に近いほど，「自動評定=見易くない」の信頼度が高くなる．この数値を一つの素性として用いる．

5.1.2 テキスト素性

大別して，以下の三種類，合計して全 10 種類の素性を用いる．

文字種ごとの文字数 漢字，平仮名，片仮名の各文字種の文字数 (全三種類).

特定の語・文字列の有無または頻度 特定の語・文字列の頻度として，「利用規約」，「相談」，「ノウハウ」，「法」[*3]，「困」[*4]の各々の語・文字列の頻度を素性として用いる (全五種類)．同様に，「例」および「Q0」，…，「Q9」の頻度の合計を素性として用いる (全一種類).

HTML タグ中の文字列パターンの頻度 HTML タグに囲まれた文字列パターンの頻度として，“>” と “<” に囲まれ

*1 https://keras.io/ja/
*2 https://github.com/keras-team/keras/blob/master/keras/applications/vgg16.py

*3 「方法」，「手法」等の語の頻度を集計することを想定し導入している．
*4 「こんな困ったことがありませんか?」等のフレーズの頻度を集計することを想定し導入している．

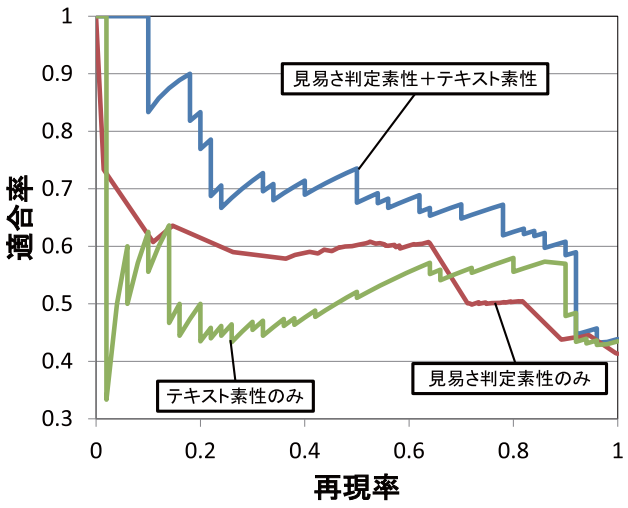

図 2: 評価結果

た一文字以上の数字列の頻度を素性として用いる (全一種類)*5.

5.2 評価手順

分類器としては，scikit-learn [Pedregosa11] パッケージにおける SVM(sklearn.SVM.SVC ツール) を用いる．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．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．そして，全体評定の値をクラスとして SVM の訓練を行う．

5.3 評価結果

素性として，(a) 深層学習による見易さ自動評定結果の素性，および，テキスト素性を併用した場合 (図 2 中の「見易さ判定素性+テキスト素性」のプロット)，(b) 深層学習による見易さ自動評定結果の素性のみを用いた場合 (図 2 中の「見易さ判定素性のみ」のプロット)，(c) テキスト素性のみを用いた場合 (図 2 中の「テキスト素性のみ」のプロット) の三通りを用いた場合を比較した結果を図 2 に示す．この結果から分かるように，深層学習によりレイアウトの見易さを自動評定した結果の素性，および，ウェブページの HTML ソーステキスト中のテキスト情報の素性という異種の素性を併用することにより，それぞれの素性を単独で用いた場合よりも全体評定の性能を改善できた．

6. 関連研究

分かりやすさの全体評定を行うのに必要な個別因子のうち，見易さ以外の因子の自動評定については，本論文の関連タスクとして，[春日 18, Han18] において，HTML 構造上の特徴を利用した見易さ以外の個別因子及び，全体評定の自動評定の手法を提案している．この研究と本稿で述べた手法を組み合わせることによって，個別因子全ての自動評定が可能になり，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できる．

この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 [石川 11, Sakai11]．また，特に，本論文における学術用語解説ウェブページの見易さの自動評定に関する関連タスクとして，[大山 17] においては，本論文とほぼ同様の深層学習手法によって，プレゼンテーションスライドの画像情報に対する印象を予測する手法を提案している．

7. おわりに

本論文では,「レイアウトの見易さ」についての因子を対象とし，深層学習を用いてこの因子を自動評定する方式 [塩川 18] によって得られる特徴量，および，用語解説ウェブページの HTML ソーステキスト中のテキスト情報の特徴量 [春日 18] を併用して，分類器として SVM を用いて全体評定の自動判定を行う手法 [春日 18] を適用し，異種の特徴量を効果的に併用可能であることを示した．

参考文献

[Han18] Han, B., Shiokawa, H., Kawaguchi, K., Utsuro, T. and Kawada, Y.: Measuring Beginner Friendliness of Chinese Web Pages explaining Academic Concepts using HTML Structures, 第 32 回人工知能学会全国大会論文集 (2018).

[石川 11] 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).

[春日 18] 春日孝秀, 塩川隼人, 韓炳材, 宇津呂武仁, 河田容英：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, 第 10 回 DEIM フォーラム論文集 (2018).

[大山 17] 大山真司, 山崎俊彦, 相澤清晴：プレゼンテーションスライドの客観評価と印象予測, 第 16 回 FIT 講演論文集, 第 3 巻, pp. 45–52 (2017).

[Pedregosa11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).

[Russakovsky14] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C. and Li, F.: ImageNet Large Scale Visual Recognition Challenge, *CoRR*, Vol. abs/1409.0575, (2014).

[Sakai11] Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K. and Lin, C.-Y.: Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection, *Proc. 4th WSDM*, pp. 187–196 (2011).

[塩川 18] 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).

[Simonyan15] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *Proc. 3rd ICLR* (2015).

*5 複数のページから構成される一連のページ群の間で任意のページに遷移するためのリンク群におけるページ数を集計することを想定し導入している．