DEIM Forum 2021 C33-3

学術用語解説ウェブページの充実度・対象読者の習熟度評定データセットの作成

Developing a Dataset of Measuring Knowledge Amount and Learning Level of Web Pages explaining Academic Concepts

大賀　悠平† 　岡田心太朗†† 　曽田　耕生† 　宇津呂武仁††† 　河田　容英††††

† 筑波大学大学院 理工情報生命学術院 システム情報工学研究群 知能機能システム学位プログラム
〒 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 知能機能システム専攻　〒 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
理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1
†††† (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に対し，解説内容の充実度，および，対象読者の習熟度を人手で評定・付与したデータセットを作成する．本データセット作成においては，ウェブページの HTML タグおよびテキスト本文の種別を考慮して，学術用語解説ウェブページ構成要素に対する情報付与を行うとともに，それらのページ構成要素に対して，主情報・例・例題等の別を定義する．そして，ページ構成要素の主情報・例・例題等の内訳の特性を主たる手がかりとして，解説内容の充実度を判定する規則，および，対象読者の習熟度を判定する規則を提案する．あわせて，本論文では，解説内容の充実度と対象読者の習熟度の間の相関を測定した結果について述べる．

キーワード　学術用語解説ウェブページ，ページ構成要素，データセット作成，充実度，習熟度

1 はじめに

近年では，様々な学術分野において初学者が学ぶ際に，インターネット上において多様な学術用語解説ウェブページが存在し，それらを利用することによって，学術用語を学ぶ際の手助けとなる場合が多い．特定の学術用語に対する学術用語解説ウェブページが複数存在する場合には，それぞれのページごとに，解説の詳しさや記載している情報の種類，想定している読者の習熟度等は様々に異なる．そのため，学習者の学術用語に対する理解度や必要としている情報の種類・量によって，最適な学術用語解説ウェブページは異なるため，各学習者は，自身の学習目的に最も適した用語解説ウェブページを探し出す必要がある．

このような背景をふまえて，本論文では，ウェブ上で学術用語を解説するページに対し，解説内容の充実度，および，対象読者の習熟度を人手で評定・付与したデータセットを作成する．本データセット作成においては，ウェブページの HTML タグおよびテキスト本文の種別を考慮して，学術用語解説ウェブページ構成要素に対する情報付与を行うとともに，それらのページ構成要素に対して，主情報・例・例題等の別を定義する．そして，ページ構成要素の主情報・例・例題等の内訳の特性を主たる手がかりとして，解説内容の充実度を判定する規則，および，対象読者の習熟度を判定する規則を提案する．あわせて，本論文では，解説内容の充実度と対象読者の習熟度の間の相関を測定した結果について述べる．

表 1　分析対象の学術用語解説ウェブページ集合

分野	サイト	用語
統計	統計 WEB	確率密度関数
	アタリマエ!	
	一番やさしい，医療統計	
	logics of blue	
	データ分析基礎知識-Albert	単回帰
	高校数学の美しい物語	確率密度関数
	AVILEN AI Trend	カイ二乗検定
	高校数学の基本問題	
解析	高校数学の美しい物語	テイラー展開
	物理のかぎしっぽ	ド・モアブルの定理
	KIT 数学ナビゲーション	テイラー展開
	新米夫婦のふたりごと	
	物理数学	
	ときわ台学	
	初等数理科学	
	未確認飛行	ラプラス変換
	竹野茂治@新潟工科大学	
	高校数学の基本問題	ド・モアブルの定理

2 分析対象の学術用語解説ウェブページ集合

文献 [11] においては，理工系学術分野を対象とし，特に，線形代数・解析・力学・電磁気・医学・IT・生物・化学・統計・プログラミングの 10 分野を対象分野として学術用語解説ウェブページを収集している．本論文では，文献 [11] において収集された学術用語解説ウェブページのうち,「統計」分野，および,「解析」分野の用語解説ウェブページに対して，サイト横断的に 18 ページを選定し，分析対象の学術用語解説ウェブページ集合とする．分析対象の学術用語解説ウェブページの一覧を表 1 に示す．

3 学術用語解説ウェブページの充実度および対象読者の習熟度

本論文では，収集した学術用語解説ウェブページに対し，表 2 に示す基準に基づいて，充実度，および，対象読者の習熟度の付与を行った．各ページに対する充実度，対象読者の習熟度の付与結果を表 4 に示す．この付与結果に基づき，充実度と対象読者の習熟度の間の相関を分析した結果を表 3 に示す．分析対象の全 18 ページ中 16 ページにおいては,「充実度=充実，対象読者の習熟度=習熟者」，もしくは,「充実度=非充実，対象読者の習熟度=初学者」という相関が成り立つ結果となった．ただし，「充実度=充実，対象読者の習熟度=初学者」となる例外的用語解説ウェブページページが 1 ページ,「充実度=非充実，対象読者の習熟度=習熟者」となる例外的用語解説ウェブページも 2 ページ観測された．

4 学術用語解説ウェブページのページ構成要素

本論文では，学術用語解説ウェブページ内の構成要素のことを「ページ構成要素」と呼び，それらを粗く「主情報」,「例・例題等」,「補助情報」に分類する．学術用語解説ウェブページの具体例における「主情報」,「例・例題等」，および,「補助情報」を図 1，および，図 2 に示す．以下の各節において,「主情報」,「例・例題等」，および,「補助情報」の詳細な定義を示す．

4.1 主 情 報

「見出し」および「見出し以外」から構成され，主として，用語そのものについての説明の記述部分を指して「主情報」と呼ぶ．

4.1.1 見 出 し

本論文では，ウェブページ内のタイトル，サブタイトル，および，ウェブページ内の文章を最も粗く分割したテキスト単位に対するタイトル部分を指して,「見出し」と呼ぶ．

4.1.2 見出し以外

本論文では，用語解説ウェブページのテキスト部分において，用語そのものについての説明の記述部分を「見出し以外」と呼ぶ.「見出し以外」は，以下に述べる「定義」,「性質」，および,「応用」から構成される．

定 義

学術用語の定義の説明テキスト部分を指す．典型的には,「~ を ~ という」等の文型が用いられる．厳格な数学的定義でない抽象的・概念的な定義を含む．当該学術用語から派生した関連用語の定義テキストも含む．ただし，当該用語の応用的概念に相当する学術用語についての説明テキストは,「応用」に相当する．例えば，当該学術用語が「確率密度関数」である場合,「連続的確率密度関数」,「離散的確率密度関数」等は関連用語とみなすが,「確率密度関数」の応用・具体例の概念に相当する「正規分布」等の定義の説明テキストは除外する．その他，具体例を交えた定義の説明テキストも「定義」に含める．

性 質

当該学術用語の性質を説明するテキスト部分を指す．例えば，当該学術用語が「確率密度関数」である場合,「総和が 1 になる」という特性の説明テキスト部分や,「確率密度関数を用いる利点」等の特徴・利用法の説明テキスト部分が該当する．当該学術用語の証明テキスト部分も「性質」として扱う．ただし，証明の途中に別の関連用語の定義テキストが挿入される場合は，その箇所は「定義」として扱う．

応 用

当該用語の応用的概念に相当する学術用語についての説明テキスト部分を指す．

4.2 例・例題等

「例」,「例題」，および,「用語解説以外のテキスト」から構成され，用語に関する情報の種類そのものを増やすことには寄与しない記述部分を指して「例・例題等」と呼ぶ．

4.2.1 例

具体的な数値を当てはめて学術用語を説明しているテキスト部分を指す．ただし，数値でなく変数を用いて説明している場合を除く．数値を当てはめた説明の途中で「このような ~ を ~ と呼びます」というテキスト部分が挿入された場合は，当該箇所は,「主情報」中の「見出し以外」(具体的には「定義」) とみなす．同様に，数値を当てはめた説明の途中で「なぜなら ~ だからです」というテキスト部分が挿入された場合は，当該箇所は,「主情報」中の「見出し以外」(具体的には「性質」) とみなす．

4.2.2 例 題

設問が独立している例題テキスト部分を指す．例題の回答・解説テキスト部分も「例題」に含める．

4.2.3 用語解説以外のテキスト

内容の理解そのものにおいては不要であるが，初学者にとっての読み易さに配慮して書かれた「コーヒーブレーク」的テキスト部分や，主情報と例・例題の間の箇所で，両者を接続させるために書かれたテキスト部分を指す．

4.3 補 助 情 報

4.3.1 ヘッダー・フッター

一つの学術用語ウェブサイト内で共通するページ構成要素であり，サイトタイトル部分の周辺に配置された UI 等が該当する．該当部分に配置されたリンク先・アンカーテキストは個々の

表 2　「充実度」·「対象読者の習熟度」の人手付与の基準

充実度	充実	定義・性質・応用について相当量の記述があり，図や式がある． (図・式は必須ではない) 指定された用語に対し，証明等の詳細な性質の記述や応用的内容が多い． 文章による抽象的な定義にとどまらず，記号・数式等を用いた数学的定義を含む． (数式で表現できない用語は除く．)
	非充実	ページ内の文章量が少なく，最低限の情報しか記述されていない． 用語に関する情報の種類そのものを増やすことには寄与せず，理解を助ける目的で記載された例・例題等の記述が多い．
対象読者の習熟度	習熟者	文体が平易でなく，不要な記述が少ない． 用語に関連する応用的な内容の記述が多い．
	初学者	文体が平易で，雑談や例・例題が多い． 数式を用いた説明が少なく，文章を用いた説明が多い． 用語を説明するための前提知識に関する説明が多い． (前提知識の説明を省略できない用語の場合は除く．)

表 3　「充実度」·「対象読者の習熟度」の間の相関

(a) 統計

		充実度	
		充実	非充実
対象読者の習熟度	習熟者	2	0
	初学者	1	5

(b) 解析

		充実度	
		充実	非充実
対象読者の習熟度	習熟者	7	2
	初学者	0	2

(c) 合計

		充実度	
		充実	非充実
対象読者の習熟度	習熟者	9	2
	初学者	1	6

表 4　学術用語解説ウェブページのデータセットにおける「充実度」·「対象読者の習熟度」·「エレメント数」

分野	サイト	充実度	対象読者の習熟度	用語	エレメント数			
					主情報 (%)	例・例題等 (%)	合計 (主情報＋例・例題等)	補助情報
統計	統計 WEB	非充実	初学者	確率密度関数	15 (37)	26 (63)	41	121
	アタリマエ!	非充実	初学者		40 (49)	42 (51)	82	152
	一番やさしい，医療統計	非充実	初学者		26 (45)	32 (55)	58	137
	logics of blue	非充実	初学者		55 (43)	74 (57)	129	552
	データ分析基礎知識-Albert	非充実	初学者	単回帰	28 (47)	32 (53)	60	198
	高校数学の美しい物語	充実	初学者	確率密度関数	28 (56)	22 (42)	50	124
	AVILEN AI Trend	充実	習熟者	カイ二乗検定	91 (97)	3 (3)	94	312
	高校数学の基本問題	充実	習熟者		1,246 (52)	1,132 (48)	2,378	336
解析	高校数学の美しい物語	非充実	初学者	テイラー展開	18 (35)	33 (65)	51	108
	物理のかぎしっぽ	非充実	習熟者	ド・モアブルの定理	9 (100)	0 (0)	9	16
	KIT 数学ナビゲーション	非充実	習熟者	テイラー展開	12 (100)	0 (0)	12	81
	新米夫婦のふたりごと	充実	習熟者		39 (68)	18 (32)	57	271
	物理数学	充実	習熟者		278 (79)	72 (21)	350	39
	ときわ台学	充実	習熟者		573 (72)	221 (28)	794	4
	初等数理科学	充実	習熟者		94 (90)	10 (10)	104	11
	未確認飛行 C	充実	習熟者	ラプラス変換	1,032 (94)	67 (6)	1,099	348
	竹野茂治@新潟工科大学	充実	習熟者		453 (99)	4 (1)	457	34
	高校数学の基本問題	充実	習熟者	ド・モアブルの定理	431 (30)	988 (70)	1,419	196

用語解説ページごとに異なる場合もある．一つの学術用語ウェブサイト内のすべてのページの末尾に特定の書籍の広告が掲載されている場合には,「広告」ではなく「ヘッダー・フッター」として扱う．

4.3.2　目　　次

当該ページ内の内容に対する目次部分が該当する．これに対して，一つの学術用語ウェブサイト内全体の目次が全てのページ共通にサイドバーに表示される場合は,「ヘッダー・フッター」として扱う．

4.3.3　他ページへのリンク

当該ページが配置された学術用語ウェブサイト内の他ページへのリンク，および，当該サイト外の他ページへのリンクが該当する．リンク前後のテキスト部分がリンク先ページについて説明している場合には，当該テキスト部分も「他ページへのリンク」として扱う．ただし，通販サイトへのリンクは「広告」として扱う．

4.3.4　コ メ ン ト

閲覧者からのコメント欄，および，当該学術用語ウェブページ著者からの返信欄が該当する．

図 1　学術用語解説ウェブページのページ構成要素「主情報」・「補助情報」の例 (「AVILEN AI Trend」サイト中の用語「固有ベクトル」解説ページの例 (https://ai-trend.jp/basic-study/linear-algebra/eigenvalue-and-eigenvector/))

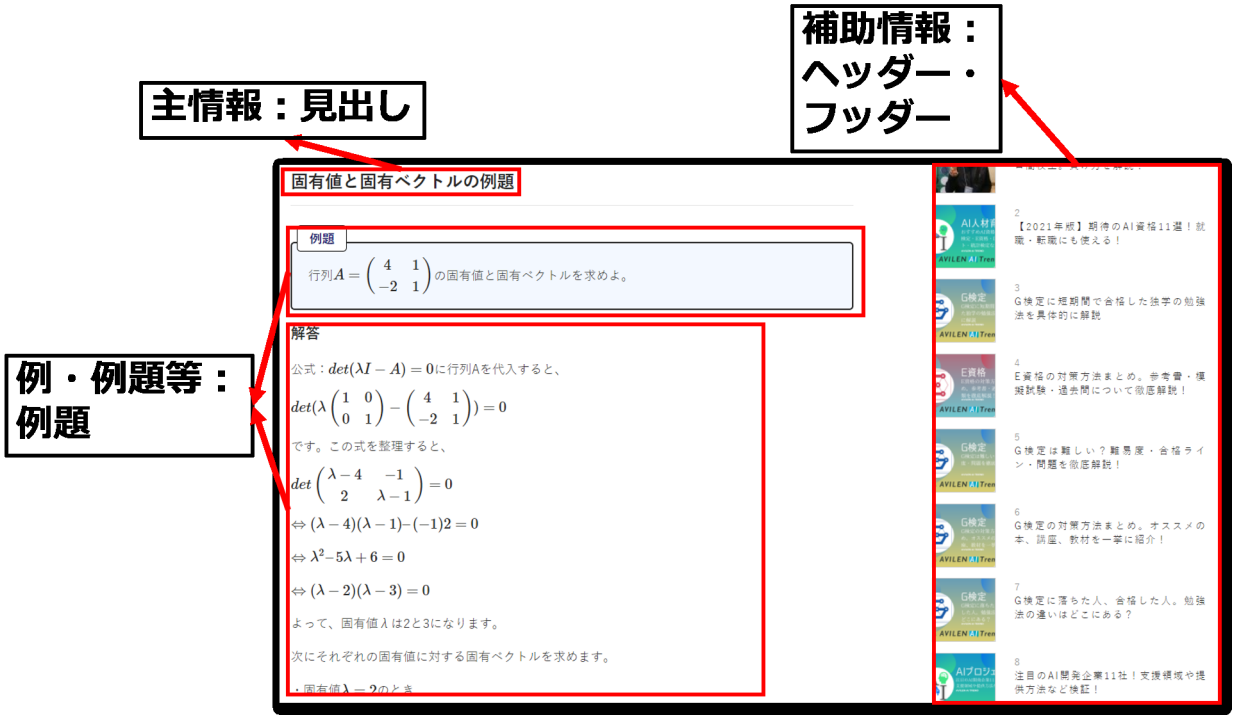


図 2　学術用語解説ウェブページのページ構成要素「主情報」・「例・例題等」・「補助情報」の例 (「AVILEN AI Trend」サイト中の用語「固有ベクトル」解説ページの例 (https://ai-trend.jp/basic-study/linear-algebra/eigenvalue-and-eigenvector/))

4.3.5　クレジット

図や表の出典を示すクレジットが該当する．当該学術用語ウェブページ著者のクレジットも「クレジット」として扱う．

4.3.6　メ　　モ

当該学術用語解説ウェブサイト・ページにおける執筆予定等の，サイト・ページ著者用のメモが該当する．コラム・注釈等

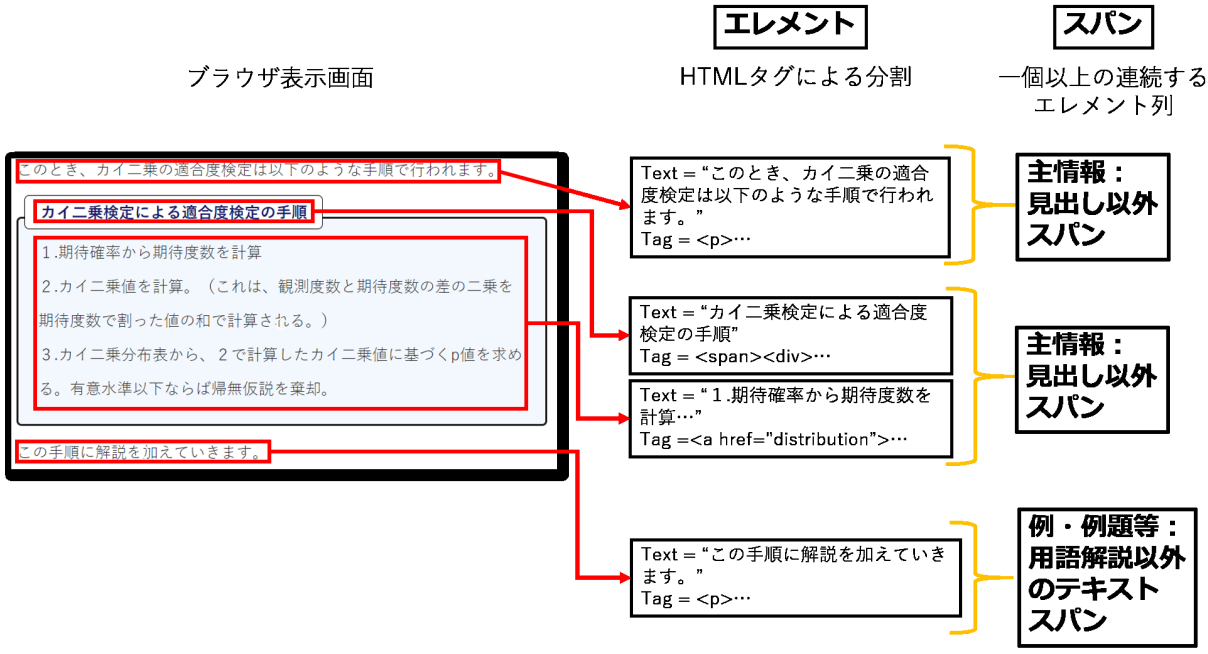


図 3　HTML タグ木構造からのエレメントの生成，および，エレメント系列へのページ構成要素スパンの付与 (「AVILEN AI Trend」サイト中の用語「カイ二乗検定」の解説ページの例 (https://ai-trend.jp/basic-study/hypothesis-testing/chi%e2%80%90square-test/))

は除外する

4.3.7　広　　告

ページ中の広告が該当する．当該学術用語解説ウェブサイト内の他ページへの誘導リンク，および，他の学術用語解説ウェブサイト内の用語解説ページへの誘導リンク等は「他ページへのリンク」として扱う．

4.3.8　HTML タグの一部

HTML のコードの一部や HTML のコメントが該当する．ブラウザで表示されない文字列すべてが該当する．

5　HTML タグ木構造上でのページ構成要素データセットの作成

本節では，図 3 に示す具体例の要領で，学術用語解説ウェブページの HTML タグ木構造を入力して，ブラウザ表示画面上のページ構成要素の範囲を人手で付与した「ページ構成要素データセット」を作成する．そのための手順として，まず，5.1 節においては，HTML タグ木構造を入力して，HTML タグ列およびアンカーテキストから構成される「エレメント」の系列を生成する手順を示す．次に，5.2 節においては，各エレメントがどの種類のページ構成要素を構成するかを同定し，一つ以上の連続するエレメントから構成されるエレメント系列に対してページ構成要素スパンを付与する手順を示す．

5.1　HTML タグ木構造からのエレメントの生成

本節では，まず，HTML タグ木構造を入力して，HTML タグ列およびアンカーテキストから構成される「エレメント」の系列を生成する．この「エレメント」は，HTML テキスト中のページ構成要素の種類を同定する際に，ページ構成要素の断片に対してその種類を判定する対象となる最小単位となるデータ構造である．一つの「エレメント」は，HTML タグ情報およびアンカーテキスト情報から構成される．「エレメント」の HTML タグ情報としては，HTML タグ木構造中の根に当たる<html>タグから葉に当たるアンカーテキストまでの各階層のタグを連結したものを用いる．例えば，<html>タグ内の<bodyl> 内の<h1>タグ内に記述されているアンカーテキストに対応する「エレメント」の HTML タグ情報は，Tag = <h1><body><html>となる．「エレメント」のアンカーテキスト情報としては，HTML タグ木構造中の一つの葉に相当する HTML タグで囲まれたアンカーテキストを用いる．

表 4 においては，本論文で分析対象とした 18 個の用語解説ウェブページに対して，人手で「エレメント」を生成した結果における「主情報」，「例・例題等」，「補助情報」の種類ごとの「エレメント」数を示す．

5.2　エレメント系列へのページ構成要素スパンの付与

前節で生成したエレメント系列に対して，本節では，各エレメントがどの種類のページ構成要素を構成するかを人手で同定し，一つ以上の連続するエレメントから構成されるエレメント系列に対してページ構成要素スパンを付与する [1]．

1：本節で付与するページ構成要素スパンは，今後，ページ構成要素スパン自動付与モデルを訓練する際の訓練事例として用いる．その際，同種のページ構成要素スパンが連続して現れる場合に，モデルによる自動付与性能を最適化するためには，どの程度の粒度でそれらを分割しておくのが適切であるかを見極める必要がある．なお，同種のページ構成要素スパンが連続する場合もそれらをあえて連結しない場合においては，一つの「主情報」スパン，あるいは，「例・例題等」ス

なお，用語解説ウェブページ内にリンク，数式，強調表現，特殊な配置の文章等が含まれる場合，それらは，HTML タグ木構造上でも特殊なタグで囲まれている．そのため，ブラウザ表示画面上は連続する一連のテキストとして表示される場合でも，エレメント系列として生成した場合には，一つの文が複数のエレメントに分割されることが起こり得る．そして，そのエレメント系列に対してページ構成要素スパンの付与を行った結果において，一文中の複数のエレメントが，二種類以上のページ構成要素スパンの系列に分割されることが起こり得る．

6 ページ構成要素の割合に基づく学術用語解説ウェブページの充実度・対象読者の習熟度判定

本節では，前節で作成した「ページ構成要素データセット」に対して，ページ構成要素の種類ごとのエレメント数とその割合，および，ページ構成要素の種類数を手がかりとして，学術用語解説ウェブページの「充実度」および「対象読者の習熟度」の判定を行う規則について述べる．表 4 において分析対象とした 18 個の用語解説ウェブページに対して，本節で述べる判定規則を適用した結果においては，分析対象の 18 ページ全てにおいて，3 節において人手で判定した「充実度」および「対象読者の習熟度」と一致する判定結果となった．

6.1 充実度判定

「ページ構成要素データセット」に対して，ページ構成要素の種類ごとのエレメント数とその割合を手がかりとして，学術用語解説ウェブページの「充実度」の判定を行う規則を表 5(a) に示す．

学術用語解説ウェブページのページ構成要素のうち，当該用語に関する主たる情報が記載されている箇所は「主情報」である．そのため，「主情報」となる定義・性質・応用の記述が多く，「例・例題等」といった理解を助けるための記述が少ない学術用語解説ウェブページは，当該用語そのものに関する記載内容が充実していると考えられる．この場合に該当し，「主情報」の割合が大きく，「充実度=充実」となる用語解説ウェブページの模式図を図 4 に示す．一方，「主情報」の割合が小さく，「充実度=非充実」となる用語解説ウェブページの模式図を図 5(a) に示す．

しかし，用語解説部分のテキストの絶対量が極めて少なく，結果としてエレメント数が極めて少ない学術用語解説ウェブページに関しては，「主情報」が占める割合が大きくても，「充実度=充実」と判定できない場合がある．「主情報」の割合が大きいもののエレメントの絶対量が極めて少ないため，「充実度=非充実」となる用語解説ウェブページの模式図を図 5(b) に示す．逆に，エレメントの絶対量が極めて多い学術用語解説ウェブページについては，「主情報」が占める割合が小さくても，結果的に十分な量の「主情報」を含んでおり，「充実度=充実」と判定できる場合も存在する．

以上をまとめると，表 5(a) に示す「充実度」判定規則となる．

パンを構成するエレメント数は約 7 個であるのに対して，一つの「補助情報」スパンを構成するエレメント数は約 1 個である．

図 4 「充実度=充実」となる学術用語解説ウェブページの模式図

本論文で分析対象とした 18 個の用語解説ウェブページに対する分析結果をふまえて，

(a) 「主情報」のエレメント数が 20 以下の場合，「主情報」の絶対量は極めて少ない．

(b) 「主情報」のエレメント数が 21~299 の場合，「主情報」の絶対量は多くも少なくもない．

(c) 「主情報」のエレメント数が 300 以上の場合，「主情報」の絶対量は極めて多い．

と判断する．そして，(a) の場合は「充実度=非充実」，(c) の場合は「充実度=充実」と判断する．一方，(b) の場合は，「主情報」＋「例・例題等」のエレメント数に対する「主情報」のエレメント数の割合が 50%以上であれば，「主情報」の割合が相対的に大きいとして「充実度=充実」と判断する．

6.2 対象読者の習熟度判定

「ページ構成要素データセット」に対して，ページ構成要素の種類ごとのエレメント数の割合，および，ページ構成要素の種類数を手がかりとして，学術用語解説ウェブページの「対象読者の習熟度」の判定を行う規則を表 5(b) に示す．

本論文では，表 5(b) の「対象読者の習熟度」判定規則においては，以下の二種類の手がかりを用いる．

(a) 「主情報」＋「例・例題等」のエレメント数に対する「主情報」のエレメント数の割合が 60%以上であれば，「主情報」の割合が相対的に大きく，「例・例題等」の割合が相対的に小さいとして，「対象読者の習熟度=習熟者」の傾向が大きいと判断する．逆の場合には，「対象読者の習熟度=初学者」の傾向が大きいと判断する．

(b) 「主情報」のうちの，4.1.2 節で述べた「見出し以外」要素として説明した「定義」，「性質」，「応用」に加え，ページ内に「図」，もしくは，「式」が含まれるか否かに着目する．具体的には，それらの「定義」，「性質」，「応用」，「図」，および，「式」の 5 つのうち，3 つ以上を含む場合には，その用語解説ウェブページにおいては，習熟者向けに詳細な記述が掲載されている可能性があるとして，「対象読者の習熟度=習熟者」の傾

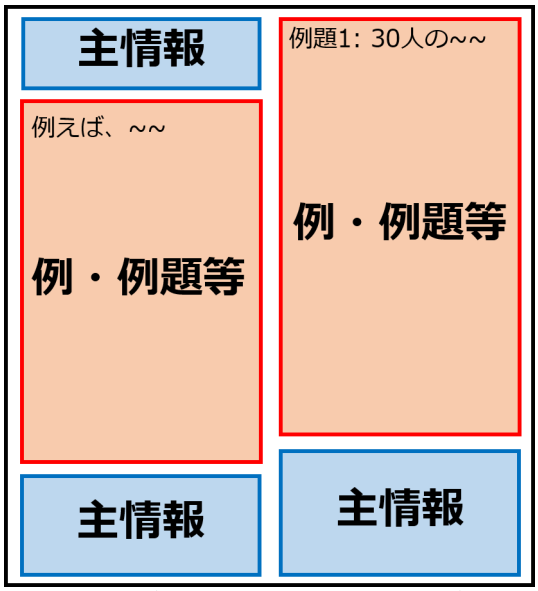
表 5　学術用語解説ウェブページの「充実度」·「対象読者の習熟度」判定規則

(a)「充実度」判定規則

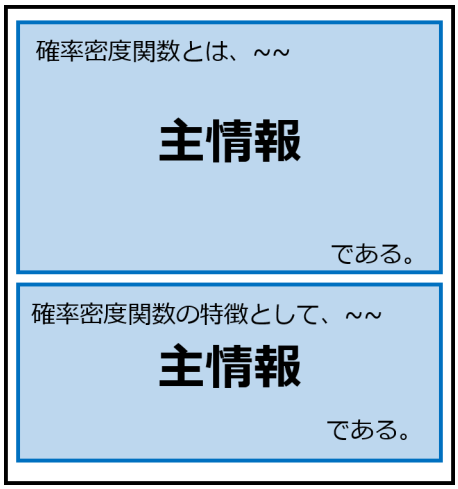
主情報のエレメント数 / (主情報＋例・例題等) のエレメント数	主情報のエレメント数 ≤ 20	21~299	≥ 300
＜ 50%	非充実		
≥ 50%		充実	

(b)「対象読者の習熟度」判定規則

主情報のエレメント数 / (主情報＋例・例題等) のエレメント数	定義・性質・応用・図・式 2種類以下が存在	3種類以上が存在
＜ 60%	初学者	初学者 習熟者
≥ 60%	習熟者	



(a) 例・例題等が多く，結果的に用語に対する情報量が少ない



(b) 簡単な定義と性質しかない

図 5　「充実度=非充実」となる用語解説ウェブページの模式図

向が大きいと判断する．逆の場合には,「対象読者の習熟度=初学者」の傾向が大きいと判断する．

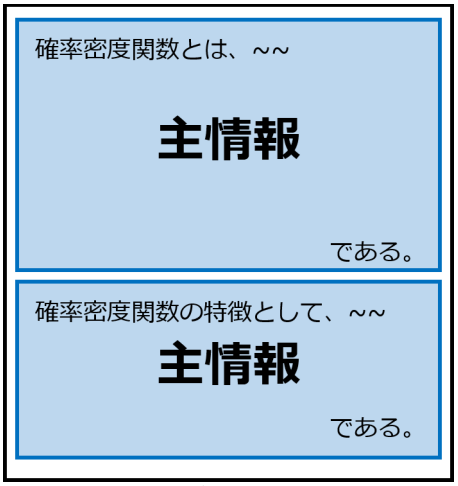
上記の (a) についての考え方として，用語の理解を助けるための具体例や例題,「コーヒーブレーク」的雑談のような文章は「例・例題等」に分類される．そのため,「例・例題等」の割合が多い学術用語解説ウェブページは「対象読者の習熟度=初学者」と考えられ，一方,「例・例題等」の割合が少ない学術用語解説ウェブページは「対象読者の習熟度=習熟者」と考えられる．ただし，上記の (b) で述べたように,「例・例題等」の割合が多くても,「定義」,「性質」,「応用」,「図」，および,「式」の 5 つのうち，3 つ以上を含む場合には，その用語解説ウェブページにおいては，習熟者向けに詳細な記述が掲載されている可能性があるとして,「対象読者の習熟度=習熟者・初学者」の両方の可能性があると判断する [2]．

2：「対象読者の習熟度」の判定においては，表 5(b) の判定規則を用いずに，表 3 における「充実度」·「対象読者の習熟度」の間の相関分析の結果に基づいて，表 5(a) の「充実度」判定規則に基づき「充実度」判定を行った後，表 3 の相関に従って「対象読者の習熟度」判定を行った方が高い判定性能が得られる．

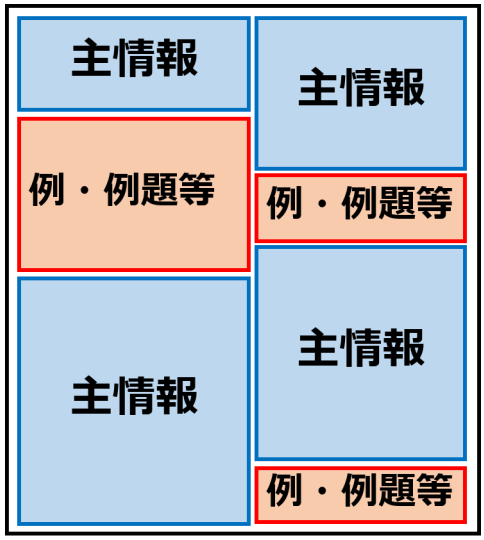
また，一般には,「充実度=充実」となる用語解説ウェブページは習熟者向け，一方,「充実度=非充実」となる用語解説ウェブページは初学者向けという傾向があるが，3 節でも述べた通り，例外も観測されている．図 6(a) は,「充実度=非充実」となるが，具体的かつかみ砕いた説明が掲載されていないため初学者には不向きであり，式だけを確認したいような習熟者向け (「対象読者の習熟度=習熟者」) の用語解説ウェブページの模式図である．一方，図 6(b) は，テキストの絶対量が極めて多いため,「例・例題等」が 40%以上を占めているが「充実度=充実」であり，かつ,「対象読者の習熟度=初学者」の用語解説ウェブページの模式図である．

7　関 連 研 究

本論文の関連タスクとして，文献 [1], [5] においては，HMTL 構造上の特徴を利用することによって，用語解説ウェブページの良否の自動評定手法を提案している．一方，文献 [6], [7] では，深層学習によって用語解説ウェブページの見易さを自動評



(a) 「充実度=非充実」だが,「対象読者の習熟度=習熟者」となる例 (表 4 中の「KIT 数学ナビゲーション」参照)



(b) 「充実度=充実」だが,「対象読者の習熟度=初学者」となる例 (表 4 中の「高校数学の美しい物語」参照)

図 6　「充実度」と「対象読者の習熟度」の相関の例外 (模式図)

定する手法，および，見易さ自動評定結果の理由を提示する方式を提案している．これらをふまえて，文献 [10] では，テキスト特徴・画像特徴を併用して用語解説ウェブページの良否を自動評定する手法を提案している．その他，文献 [2], [3] においては,「分かり易さ」と「見易さ」のうち,「分かり易さ」のみが充足され,「見易さ」が充足されない場合，および，逆に,「見易さ」のみが充足され,「分かり易さ」が充足されない場合に焦点を当て，それぞれ「見易さ」あるいは「分かり易さ」を損なう因子群を網羅的に分析している．以上の先行研究の中での本論文の位置付けとして，本論文のデータセットを教師データとして，用語解説ウェブページの充実度，および，対象読者の習熟度を自動判定することにより，個々の学習者の利用目的に応じて用語解説ウェブページを選択的に推薦することが可能になる．

その他，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 [4], [9]．その他，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，文献 [8] においては，文献 [6], [7] 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．

8　おわりに

本論文では，文献 [11] において収集された理工系学術分野 10 分野の学術用語解説ウェブページのうち,「統計」分野，および,「解析」分野の用語解説ウェブページを対象として，ページ構成要素に着目して記述の内容の充実度と対象読者の習熟度を評定するためのデータセットを作成した．今後の課題として，本論文で作成したデータセットの仕様のもとで，文献 [11] で収集・分析対象となった理工系学術分野全 10 分野を対象に学術用語解説ウェブページを収集し，その充実度・対象読者の習熟度の情報付与を行うことが挙げられる．これにより，本論文の仕様の適用可能性の検証を行う．さらに，それらのデータセットを教師データとして，本論文において情報付与の対象としたページ構成要素，充実度，対象読者の習熟度を自動同定するモデルの訓練・評価を行う．

文　献

[1] B. Han, H. Shiokawa, K. Kawaguchi, T. Utsuro, and Y. Kawada. 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.

[2] 廣花智遥, 岡田心太朗, 塩川隼人, 韓炳材, 宇津呂武仁, 河田容英, 神門典子. 学術用語解説ウェブページの分かり易さ・見易さ因子分析および見易さ自動評定結果の理由提示. 情報処理学会研究報告, Vol. 2019–IFAT–134/2019–DC–112, No. 8, pp. 1–8, 2019.

[3] 廣花智遥, 岡田心太朗, 宇津呂武仁, 河田容英, 神門典子. 学術用語解説ウェブページの良否評定のための分かり易さ・見易さ因子の分析. 第 33 回人工知能学会全国大会論文集, 2019.

[4] 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.

[5] 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.

[6] 岡田心太朗, 塩川隼人, 韓炳材, 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 深層学習による学術用語解説ウェブページの見易さ自動評定結果の理由提示. 第 11 回 DEIM フォーラム論文集, 2019.

[7] 岡田心太朗, 曽田耕生, 大賀悠平, 宇津呂武仁, 河田容英. 学術用語解説ウェブページにおけるページ構成要素を考慮した見易さの評定および理由提示. 第 13 回 DEIM フォーラム論文集, 2021.

[8] 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.

[9] 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.

[10] 塩川隼人, 岡田心太朗, 韓炳材, 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 深層学習を用いた学術用語解説ウェブページの分かり易さ・見易さの自動評定. 第 11 回 DEIM フォーラム論文集, 2019.

[11] 曽田耕生, 大川遥平, 岡田心太朗, , 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 学術用語解説ウェブページ見易さ評定モデルのサイト単位適用事例の分析. 第 12 回 DEIM フォーラム論文集, 2020.