原著論文 Original Paper

レビュー情報を用いた学術本の難易度推定

An Estimation of Difficulty for Academic Books using Reviews

中山 祐輝 Yuki Nakayama
金沢大学大学院自然科学研究科
Kanazawa University Graduate School of Natural Science & Technology
6174naka@blitz.ec.t.kanazawa-u.ac.jp

南保 英孝 Hidetaka Nambo
（同　上）
nambo@blitz.ec.t.kanazawa-u.ac.jp

木村 春彦 Haruhiko Kimura
（同　上）
kimura@blitz.ec.t.kanazawa-u.ac.jp

keywords: information retrieval, difficulty, learning support, recommender system

Summary

A collaborate filtering has been generally used as a method which recommends items to customers. However, recommending academic books, it need to consider difficulty of them and individual amount of knowledge as well as user's preference. If the recommendation method considers only user's preference, they might regret after buying or reading recommended book because it won't match user's appropriate level. In this paper, we focus on academic books and propose a method which estimates the difficulty of academic books using user's reviews. Estimating difficulty of books will support users to search and recommend academic books that match user's skill. Moreover, we evaluated applying our method to academic text books about C programming Language. We verified that our method is more effective than traditional methods for academic books.

1. はじめに

ある分野の専門知識を深めたい時に考えられる方法として，本を読むことや，検索エンジン，e-learning を利用することなどが挙げられる．しかし，e-learning は情報量が本や検索エンジンと比べて少ないことや，誰もが e-learning を利用できる学習環境に必ずしも恵まれていないという欠点がある．また，検索エンジンは短時間で本よりも膨大な情報を獲得できるという利点があるが，情報が多すぎるためにどの情報が一番有用なのかが分からず，情報の信憑性の問題もある．一方，本は情報量が多く，かつ信憑性も高く，詳しい解説が載っているため入門者から上級者まで柔軟に知識を得ることができる．しかし，レベル・ジャンルが多岐にわたるため，どの本を選択すれば良いかわからないという問題がある．特に学術的な専門書 (以下，本論文では学術本と呼ぶ) を読むときにはこの問題が顕著に表れる．この問題に対し，Amazon などでは本を選択する時の指針として，協調フィルタリングを用いた推薦システムが広く用いられている．これは過去の閲覧・購入履歴などの嗜好情報に基づくものである．しかしながら，嗜好情報だけで本を選択しても，その本が自分の知識やレベルに合っている本であるとは限らず，結果として，時間やお金の浪費につながることも考えられる．また，Amazon などでは検索結果を関連度・人気度・価格などのキーで並び替えることが出来る．しかしながら，学術本を一覧表示させるときには難易度をキーにすると有用であると考えられる．難易度で並び替えることにより，初学者にとってはスタートとなる本が見つけやすくなることが期待できる．また，ある本を十分理解できたユーザーにはより高いレベルの本を選択するとき，逆に挫折してしまったときには少しレベルの下がった本を選択するときの指標となる．つまり，学術本を読む際には，個人の知識量や本の難易度を考慮して本を選択することが重要である．そこで，本論文では本の難易度に焦点を絞り，学術本に難易度を付与する手法を提案する．そして，本の難易度を個人の知識量推定の 1 つの指標にできると考え，知識量推定につなげていくことを目的としている．以下，第 2 章で関連研究，第 3 章で提案する本の難易度を推定する手法について述べる．第 4 章では難易度の算出の妥当性を検証した実験・評価方法について説明し，第 5 章でその実験結果および考察を示す．最後に，第 6 章で本論文のまとめとする．

2. 関連研究

2·1 レビュー情報を用いた情報推薦と検索

情報検索および推薦にレビュー情報を用いる研究はいくつか行われている．[中谷 08] ではマーケティング分野

で重視されている sence, feel, think, act, relate の 5 つの観点からレビューを用いてゲームを分類し，ユーザーの経験的価値に合ったゲームを推薦している．レビューを用いて情報を可視化する点では本研究と似ているが，学術本ではその本が理解出来るかが重要なため，5 つの観点から得られる経験的価値は利用できない．学術本における経験的価値とはその本が難しかったか，易しかったかであると考えられる．また本研究では本を対象としている点も異なっている．

[鈴木 09] ではユーザーが好むコンテンツのタイトルからレビューが類似しているコンテンツをユーザーに推薦している．これは，本研究と同じく本を扱っているが，学術本ではなく小説を扱っているという点で異なる．また，固定レビューと変動レビューに分類しているが固定レビューを決めるための固有キーワードを登場人物の名前や地名としているので学術本に関するレビューには適用できない．

[倉島 07] では比較評価情報として評価対象，比較対象，属性，評価を抽出し，対象間の優劣をランキングで可視化している．対象を映画としており，評価として「良い」「優れている」という比較の方法をしている．本研究では優劣ではなく難易度で比較しようとしている点で異なる．

[杉木 08] では自然言語で記述された検索クエリに適応する商品検索手法を提案している．学術本であれば「ユーザーのレベルに合った本」などと検索できればよいが，自然文のクエリ以外に個人の知識量も考慮して検索する必要があるため，この手法では難しいと思われる．

2·2 テキスト情報を用いた難易度推定

テキストから対象の難易度を推定する研究はいくつか行われている．[Nishihara 05] では各 Web ページの特徴語 (入力キーワードの説明に用いられている単語) を抽出し，特徴語の難しさの評価値の和から Web ページの難易度を算出している．特徴語の難しさとして出現頻度が低いほど，難しく意味を捉えにくいと考え，特徴語の出現頻度を用いている． [Nishihara 05] は被験者に付けてもらった Web ページの難易度と提案手法の Web ページの難易度との順位相関を評価としており，中程度の相関が得られているが難易度を推定できるまでには至っていない．

[Nakatani 09] は文書中の専門用語の出現頻度と，文書の読みやすさの指標であるリーダビリティ [Sato 08] から，一般的なユーザーにとっての理解の容易さを文書ごとに測定し，測定結果を理解容易度として Web ページに与えている．リーダビリティの指標は小学生 (1-6)，中学生 (7-9)，高校生 (10-12)，高校生より高いレベル (13) の 13 段階に区分されている．[Nakatani 09] を本に適用するためには，専門用語の頻度に加え，難しさも考慮する必要がある．なぜならば学術本には専門用語がつきものであり，どの本にも同じぐらい出現すると考えられるからである．また，Web ページは不特定多数の人が書いているのに対し，本は推敲や校正が行われているため，リーダビリティは高くなると予想できる．すなわちこの手法では本の難易度を推定することは難しいと思われる．

[中條 04] は英語テキストおよび日本語テキストに難易度を付けている．難易度の指標として英語テキストは 1) 英語リーダビリティ公式によるリーダビリティ・スコア，2) 日本の中学・高等学校英語教科書に出現しない語の割合を用いている．また，日本語テキスト部分については，1) 日本語能力試験の語彙 1,2 級の割合，2) 漢字含有率，というテキストの内容理解に直接影響する要因を測る指標を用いている．[中條 04] は英語・国語のような一般向けの本には適用できるが専門用語の入った専門書などには適用できない．

2·3 読者ネットワークを用いた学術本の難易度推定手法

[三好 10] は学術本を主眼として本の難易度を推定している．以下の 3 つの仮説を基に誰がどの本を読んだかという関係から成る読者ネットワークを用いて難易度の推定を試みている．(1) ある分野の本を多く読んでいる人は，その分野について詳しい．(2) 詳しい人に読まれている本ほど難しく，詳しくない人に読まれている本ほど易しい．(3) 難しい本を読んでいる人は詳しく，易しい本を読んでいる人は詳しくない．[三好 10] は一般書・専門書にも対応しており，本論文ではこの手法を従来手法とし，比較を行う．従来手法のアルゴリズムは以下の通りとなっている．従来手法は上記の 3 つの仮説に基づき，Web ページの重要度の指標となる HITS[M.Kleinberg 98] アルゴリズム内の authority と hub の概念を用いている．図 1 に示すように本と読者のそれぞれのノードが完全に分割された 2 部グラフを扱う．エッジは読者と読者が読んだ本と結ぶ読者関係を表し，無向グラフである．この手法では，「本」ノードは尺度として難易度をもち,「読者」ノードは習熟度をもつ．難易度は HITS における authority に相当し，習熟度は hub に相当する．読者ネットワークが与えられた時，HITS と同様に，反復計算により読者 u の習熟度 a_u, 本 i の難易度 d_i を次式で求めることができる．

$$a_u = \sum_{i \in I_u} (\dot{d}_i - p_a) \tag{1}$$

$$d_i = \sum_{u \in U_i} (min(\dot{a_u}, p_d) - 0.5) \tag{2}$$

$\dot{a_u}$ は a_u を正規化したもの，$\dot{d}_i$ は d_i を正規化したもの，I_u は読者 u が読んだ本の集合，U_i は本 i を読んだ読者の集合である．正規化の方法は偏差値を 100 で割っている．p_a，p_d はパラメータである．初期値として全ての本の難易度を 0.5 とし，計算結果が収束するまで式 1,2 を繰り返す．[三好 10] では読者ネットワークを構築した時点で読者が未読であるのか，読み始めたばかりなのか，一通り

読んだのかなどといった進度情報を考慮していない．例えば，未読であればユーザーの習熟度は皆無であり，一通り読んでいれば習熟度は上がっている．よって，本の難易度に誤差が発生すると考えられる．

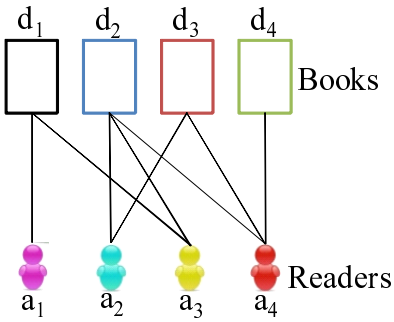


図 1 従来手法の読者ネットワーク例

3. 提 案 手 法

3·1 提案手法の概要

本論文における難易度とは，効率良く学習するための段階的に示した学習過程 (順番) と対応付けた値と定義する．つまり，難易度が相対的に高いほど後半に学習すべき本ということになる．このように定義される難易度の妥当性については 4·1 節で後述する．本論文で提案する難易度推定手法は，[三好 10] とはアプローチが異なり，自然言語処理の技術を用いたものとなっている．具体的には，Amazon の本のレビューを用いて相対的な本の難易度を推定する．レビューはユーザーがある物に対する興味度合いや役に立ったかについて記述されたテキストである．レビューデータを用いる利点としては専門的な本であればその本が難しかったのか，読み易かったのかなどといった評価がされており，難易度推定のためのデータとして期待できる．また，レビューはユーザーが本を読んでいることが前提であり，読者の進度の問題はない．そこで，レビューの特性について，我々は次のような 2 つの仮説を立てる．

仮説 1 難易度が低い本ほどレビュー内で「わかりやすい」や「読みやすい」などの評価を受けている

このような評価をしているユーザーが多いほどその本は難易度が低いと考えられる．実際にレビュー内に書き込まれている評価を図 2 に示す．本論文では図 2 のような下線部の形容詞の評価を評価表現と表記する．

仮説 2 (1) 難易度が高い本ほどレビュー内にその本が関連する分野の専門用語がより含まれる傾向にある (2) また含まれる専門用語が難しくなる傾向にある

難易度の高い本は，難易度の低い本に比べ難しい専門用語が登場するため理解するのがより難しくなる．ユーザーはその本の内容についてレビューを書き込むのでレビュー内にそのような専門用語が出現すると考えられる．それに伴い，レビュー内の専門用語の出現頻度は高くなると考えられる．

・また三色刷で 見やすく 図も挿入されていて わかりやすい.
・その中で、最も 理解しやすかった のがこの本です.
・又、初心者に教えるときに補助教材にしながら説明をしても とっつき易く て、いいかもしれない.

図 2 レビュー対象が本の評価例

提案手法は以上の仮説に基づきレビュー内に出現する評価表現と専門用語を用いて C 言語に関する本を対象とし，学術本の難易度の推定を試みる．提案手法の流れを図 3 に示す．提案手法は統計処理と計算処理の二つに分かれている．以下，3·2 節～3·4 節で統計処理について，3·5 節～3·6 節で計算処理について説明する．

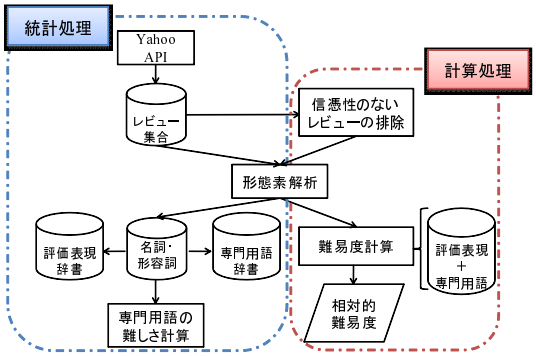


図 3 提案手法の流れ

3·2 評価表現辞書の構築

まず，Yahoo!が提供している Yahoo!API [*1]を利用し，Amazon からレビューを取得する．本論文では平成 23 年 3 月 17 日より過去である全てのレビューを用いる．そして評価表現および専門用語を抽出するために MeCab[*2]を用いて形態素解析を行う．なお，レビューを形態素解析するためには，レビューを文単位に分割する必要がある．レビューの分割には文末記号 (。. !?) を用いた．その際，文末にでてくる括弧の補足情報，鍵括弧や括弧の中の文末記号など，本来分割すべきではないところでは経験則により手動で行った．次に，形態素解析によって得られた品詞情報から「分かりやすい」・「読みやすい」といった形容詞を抜き出し，評価表現辞書として登録する．評価表現辞書は表 1 のように 2 つのジャンルを合わせた 5113 文で高頻度で出現した表現を MeCab 辞書に登録する．辞書中の評価表現の一覧を表 2(a) に示す．

評価表現辞書には「易しさ」を表現する語のみを登録しているが，「わかりにくい」や「読みにくい」など「難しさ」を表現する語もレビュー中には出現する．よって，両方の表現を手掛かりにして難易度を推定していく方法も考えられる．レビュー中には図 4(a)，(b) の下線部のよ

*1 http://developer.yahoo.co.jp/webapi/search/
*2 http://mecab.sourceforge.net/

うに「易しさ」と「難しさ」を表現する形容詞においてレビュー対象の本を評価していない語が見られる．また，図 4(b) の 1 番目の例では「難しい」という表現において否定語「ない」が付随する語も見られる．これらの語はノイズと考えられ，文脈を解析しないとノイズを除去することは難しい．表 2(a)，(b) はそれぞれの表現におけるノイズ発生率を示している．この表から「難しい」表現のノイズ発生率の方が高く，また出現頻度が低いため，本論文では「難しさ」を表現する形容詞は扱わないこととした．

・微分積分に関してはこれよりも 分かりやすく て良い本が山のように存在する。
・今まで私が見た本は、本来難しいことを わかり易く しようとするがあまり、逆に考え込むと深みにはまるような傾向がありました。

(a)「易しさ」を意味する評価表現

・そのほとんどが，中級レベル以上向けに書かれたような本が多く丸っきりの初心者には とっつきにくい 感じがありました．この本は，そんなに 難しい ことは書いて いません．
・説明がシンプルで 分かりにくい 概念をあっさり説明していた

(b)「難しさ」を意味する評価表現

図 4 レビュー対象本の内容を評価していないまたは否定語「ない」を含む評価表現例

表 1 コーパスの概要

分野	冊数	レビュー数	レビューの文の数
C 言語	59	445	3247
解析学	55	237	1866

表 2 評価表現のノイズ発生率

(a)「易しい」表現

評価表現	頻度	ノイズ頻度
わかりやすい	166	24
分かりやすい	106	13
読みやすい	51	8
分かり易い	28	1
見やすい	27	3
理解しやすい	23	1
わかり易い	9	1
解りやすい	7	0
とっつきやすい	6	2
読み易い	4	1
理解し易い	3	0
見易い	3	0
とっつき易い	3	0
総頻度	436	54
ノイズ発生率	**12.39%**	

(b)「難しい」表現

評価表現	頻度	ノイズ頻度
難しい	123	66
わかりにくい	16	6
分かりにくい	15	7
読みにくい	7	2
理解しにくい	5	4
とっつきにくい	8	5
理解しづらい	2	2
分かりづらい	4	0
総頻度	182	92
ノイズ発生率	**50.55%**	

3·3 専門用語辞書の構築

本論文では専門用語取得の自動獲得を主眼としていないので，手動による専門用語辞書の構築を行う．まず，専門用語は単名詞の他に,「ポインタ変数」といった複合名詞から成るものもある．そこで，専門用語抽出には N-gram モデルを用い，コーパスを C 言語の 3247 文とし，uni-gram,bi-gram,tri-gram に対して統計をとった．統計から得られた単名詞・複合名詞のうち出現頻度が 2 以上でかつ新 ANSIC 言語辞典 [平林 97] に載っている用語を専門用語として扱うこととした．結果として 161 語の専門用語が得られた．抽出された専門用語の一覧を付録 A(() 内の数字は頻度) に示す．

3·4 共起に基づいた専門用語の難しさ

本論文では専門用語の出現頻度に加え，難しさを考慮し難易度を推定する．難しい専門用語とは，難しく意味を捉えにくい用語のことである．また，一般的に専門用語は出現頻度 (以下 tf 値と書く) が低いほど難しい傾向がある [中川 03]．しかし,「ポインタ」について詳しく書かれた本のレビュー内には「ポインタ」という語が頻出する．そのため，付録 A を見ると専門用語「ポインタ」の tf 値が高くなっている．しかし，C 言語ではポインタは意味を捉えにくく，難しい用語である．一方，専門用語「16 進数」は情報学の基礎となる用語であるが，tf 値が小さく，難しい用語と判断されてしまう．つまり，tf 値を専門用語の難しさとすると，本来の用語の難しさと乖離してしまう専門用語が含まれていることになる．本論文ではこのような語をノイズと呼ぶ．さらにレビュー中に出現する専門用語の tf 値はかなり差があることもわかる．よって，tf 値を本の難易度に取り入れると「C 言語」のような tf 値が極端に高い用語に依存してしまうことになる．つまり，ノイズや頻度の大差が難易度の計算に影響を及ぼすと考えられる．そこで，本論文では単語の共起性に着目し，各用語の tf 値を大小関係を考慮した形で変化させることを考える．共起性は，専門用語の関連性を知る目安として使うことができる [相澤 00, 伊藤 07]．また共起関係にあるものを 1 つにまとめた時，それらは一つの概念を表しやすい [松尾 02]．図 5 は 3 冊の本について実際に書き込まれているレビュー各 1 文を載せたもので，1 文中に共起している専門用語 (アンダーライン) の例を示している．このように，共起して出現する専門用語は同程度に意味が捉えにくいと考えられる．例えば，ある用語がノイズであっても，関連した他の用語と難しさをまとめることにより，本来 tf 値が持っているノイズを抑制でき，かつ tf 値の大差を軽減できる．このように tf 値の傾向を変化させることで，本の難易度との関連性がより高くなると考えた．本論文では，共起性の尺度として相互情報量[*3]を用い，共起に基づいたクラスタリングを行うことで tf 値を変化させる．共起頻度は文単位で得るものとし，以下のアルゴリズムにより共起クラスタリングおよびクラスタ毎の難易度を付与する．

準備：クラスタ集合を C とし，$|C| = 0$ とする．専門用語 $w1$,$w2$ の共起頻度を要素とする 161 × 161 の共起行

*3 相互情報量は以下で表される．
$MI(w1, w2) = \frac{Nfreq(w1,w2)}{freq(w1)freq(w2)}$

・C 言語 特有の機能 (ポインタ や 構造体 等) を説明に重点を置いた、いわゆる「2 冊目本」を探していました.
・printf 関数 や scanf 関数getchar関数 などそして、四則演算やそれを扱うための注意点などもかかれています。
・そもそもがプリプロセス処理をやってから本チャンのコンパイル をし、結果を アセンブラ で出し、オブジェクトファイル にし、リンカー で ライブラリ とくっつける…

図 5 レビュー中の専門用語共起の例

列を生成し，以下のステップで処理を行う.
ステップ 1：専門用語 $w1$ と $w2$ のペア[*4]について，相互情報量 $MI(w1,w2) > \alpha$ のとき，以下の (1)〜(4) のうち，条件を満たすものを実行する.
(1) クラスタ集合 C の要素である全てのクラスタ $C_i(i,j:0 \sim |C|)$ について，$w1$，$w2$ がいずれも C_i の要素でなければまたは $|C|=0$ のとき，新規クラスタ $C_{|C|}$ を作成し，クラスタ集合 C の要素とする．また，$w1$,$w2$ を $C_{|C|}$ の要素とする.
(2)$w1$，$w2$ がそれぞれクラスタ集合のいずれかののクラスタ C_i, C_j に属し，$i \neq j$ であれば，2 つのクラスタを統合し 1 つにまとめる．クラスタ番号を整理する.
(3)$w1$ がクラスタ集合のいずれかのクラスタ C_i に属し，$w2$ がどのクラスタにも属さない場合，$w2$ を C_i の要素とする.
(4)$w2$ がクラスタ集合のいずれかのクラスタ C_j に属し，$w1$ がどのクラスタにも属さない場合，$w1$ を C_j の要素とする.
全ての組み合わせが終わるまでステップ 1 を繰り返す.
ステップ 2：各クラスタの難易度をクラスタ内の tf 値の中央値とする.
ステップ 3：専門用語の難しさ $Diff_{tech}(w)$ を式 3 で決定する.

$$Diff_{tech}(w) = \begin{cases} Diff(C_i) & \text{if } w \in C_i \\ tf(w) & \text{otherwise} \end{cases} \quad (3)$$

ある専門用語 w がいずれかのクラスタに属していればクラスタの難易度をその専門用語の難しさとする．対して，どのクラスタにも属さない専門用語は元々の tf 値を用語の難しさとする．専門用語の難しさ $Diff_{tech}(w)$ は tf 値に基づくため，値が小さいほど難しい用語と判断される.

3·5 信憑性の低いレビューの排除

レビューの中には宣伝，誹謗中傷，少数意見など低品質と考えられるものが含まれている．そのようなレビューは難易度の計算に影響を及ぼすと考え，レビューの信憑性を考慮する．ここでは参考度 f[小倉 08] に基づいて低品質と考えられるレビューを排除した．参考度 f は次式で定義される.

$$f = \frac{Y}{N} \quad (4)$$

N はあるレビューに対する総投票数を表し，Y は「そのレビューに同意する」に投票した数を表す．式 4 に示す値が高ければ，多くのユーザーから同意を得ていることを示し，レビューの信頼度が高いということを表している．[小倉 08] では信憑性の低いレビューは $f \leq 0.3$ であるとしている．また，本における参考度のヒストグラムを求めたところ，$f = 0.5$ あたりの出現頻度が比較的多くなったという結果が得られた．この結果から $f \leq 0.5$ であるレビューを除いてしまうと，1 冊あたりのレビュー数が激減し，難易度計算に用いるレビューデータが疎になってしまう．また，参考度が高くても投票数 N が少ないレビューの信憑性は未知数である．よって本論文ではレビューの信憑性とデータ量を考慮し，$f \geq 0.4$ かつ $N \geq 5$ であるレビュー $R_{i,f}$ を難易度の計算に用いることとした.

3·6 難易度の計算

提案手法では 3·2 節で挙げた評価表現，専門用語の出現頻度，3·4 節で挙げた専門用語の難しさの 3 つの指標を組み合わせて難易度を計算する．まず本 i の 1 文の形態素集合 sen を要素とするレビュー集合を $R_{i,f}$, 評価表現辞書内の表現集合を W_{easy} とし，評価表現度 $Easy(i)$ を次式で定義する.

$$Easy(i) = \frac{1}{|R_{i,f}|} \sum_{sen \in R_{i,f}} n_e(w, sen) \quad (5)$$
$$\text{ただし}, w \in W_{easy}$$

$n_e(w, sen)$ は形態素集合 sen 中に存在する評価表現の個数を表す．式 5 は評価表現が多く含まれているほど高くなる値であり，仮説 1 によると易しい本ほど大きくなる指標である.

次に本 i の専門度 $Tech(i)$ を次式で定義する.

$$Tech(i) = |R_{i,f}| \frac{\sum_{sen \in R_{i,f}} \sum_{w \in sen} Diff_{tech}(w)}{(\sum_{sen \in R_{i,f}} n_t(w, sen))^2} \quad (6)$$
$$\text{ただし}, w \in W_{tech} \text{かつ } w \notin w_{parent}$$

W_{tech} は 3·3 節で構築した専門用語辞書の集合，w は専門用語辞書に含まれている用語である．$n_t(w, sen)$ は形態素集合 sen 中に存在する専門用語の個数を表す．式 6 では本 i のレビュー $R_{i,f}$ 内に含まれる専門用語に対して 3·4 節の方法で専門用語の難しさ $Diff_{tech}(w)$ を求め，それを平均したものを本 i のレビュー 1 行あたりに出現する専門用語の個数で割っている．仮説 2(2) に基づけば 3.4 節で求めた $Diff_{tech}(w)$ は本が難しいほど値は小さくなる．また，仮説 2(1) によると，レビュー 1 行あたりに出現する専門用語の個数は本が難しいほど大きくな

*4 $w1$ と $w2$ の組み合わせは ${}_{161}C_2 = 12880$ 通り

る．よって，両者を割った $Tech(i)$ は難しい本ほど値が小さくなる．両者を割るのは専門用語の難しさと出現頻度のいずれかの指標が異常値を示したときに，他方の指標で補正を行うためである．提案手法ではレビュー中から得た名詞から専門用語を抽出するために N-gram モデルを用いているため，次のような問題が発生する．例えば，bi-gram の語である「ポインタ-変数」が十分な出現頻度で，専門用語辞書に登録されている場合，その用語の一部分からなる「ポインタ」や「変数」も難しさの指標としてみなしてしまう．これを防ぐために，レビュー 1 行で「親となる用語 w_{parent} を持たない用語 w」だけを難易度計算に用いるという制約を設けた [藤村 05].

最終的に式 5 と式 6 との和を本 i の難易度 $Diff(i)$ とする．

$$Diff(i) = 1 - (Easy(i) + Const \times Tech(i)) \quad (7)$$

式 7 の値が大きいほど難易度が高くなる．$Easy(i)$, $Tech(i)$ のどちらかの指標に仮説に反するような異常値が生じれば，他方の正常値を足すことでその異常値を緩和する方向に働く．また，両者とも仮説に基づく正常値であれば精度がよくなることが期待できる．なお，$Const$ は正の定数である．

4. 評 価 実 験

4·1 実 験 方 法

実験 1：提案手法の精度を検証するために，被験者に本を見てもらい，難しい順に並び替えてもらう．ジャンルは C 言語の本を (A)〜(I) の 9 冊 (表 3) とし，被験者として C 言語の知識がある高専の情報系 4 年生の学生 17 人を対象にした．被験者には C 言語の学習カリキュラムに基づき，低学年で実施する内容が多く載っているほど易しく，高学年で実施する内容が多く書かれているほど難しいと判断してもらった．同程度の難易度と判断した場合には見やすさ[*5]を考慮してもらった．最も難しいものは 1 点，最も易しいものを 9 点とスコア付けしてもらい 17 人分の平均値を正解データとした (正解データ 1 とする)．学習カリキュラムは教員がある分野を学習しやすいように決めた学習過程である．教員が決めた学習過程にしたがって被験者が評価しているので正解データ 1 は効率の良い学習過程を段階的に示したデータとなる．このような難易度の定義により，ユーザーが学習をするときには教員が立てたカリキュラムにしたがって学習でき，効率良く学習できると考えられる．なお，信憑性の低いレビューを排除するために参考度を適用し，適用後のレビュー数が 5 未満の本、用語辞典，解答本，演習本は難易度の計算には無理があると考えたため除外した．従来手法は [三好 10] を参考にメディアマーカー[*6]に登録されている情報を利用し，読者ネットワークを構築した．

実験 2：仮説 2(2) を検証するために情報系の大学生 17 人に専門用語 161 語に対してアンケートを実施した．アンケートの内容はある用語を「1. 意味は分からない，聞いたこともない．2. 聞いたことがあるが説明はできない．3. 聞いたことがあり，どういう場面で使うかまたは意味も知っている.」の 3 つに振り分けるものである．それぞれの選択肢のスコアを 1 点，2 点，3 点とし，17 人分の平均値を用語の難しさとした (正解データ 2 とする).

実験 3：ある本の改訂版が出版されていれば，それらは同程度の難易度になるはずである．そこで，本 (C)・(D) の改訂版である『やさしい C 第 3 版 (本 (J) と表記する)』と『猫でもわかる C 言語プログラミング第 2 版 (本 (K) と表記する)』を 9 冊の本の中に追加し，(J)・(K) の難易度を提案手法で計算し，計 11 冊の難易度と正解データとの順位相関を求める．改訂後の正解データは改訂前と同じ難易度に設定した．改訂後を混在させた 11 冊の難易度データを正解データ 3 とする．

表 3 実験に用いる本 11 冊

記号	本タイトル	全レビュー数	参考度適用後レビュー数
(A)	プログラミング言語 ANSI 準拠第 2 版	43	32
(B)	新 C 言語入門 ビギナー編	10	5
(C)	やさしい C 第 2 版	17	14
(D)	猫でもわかる C 言語プログラミング	16	15
(E)	新 C 言語入門 シニア編	10	7
(F)	エキスパート C プログラミング	8	6
(G)	明解 C 言語入門編	41	23
(H)	新 C 言語入門 スーパービギナー編	15	6
(I)	C 言語ポインタ完全制覇	23	19
(J)	やさしい C 第 3 版	13	5
(K)	猫でもわかる C 言語プログラミング第 2 版	10	7

4·2 評 価 方 法

実験 1 と 3 については，被験者によってつけられた本 9 冊の難易度の正解データ 1 の大小関係と，提案手法によってつけられた大小関係との順位相関を評価とする．実験 2 では正解データ 1 と，正解データ 2 から計算される「レビュー中に出現する専門用語 1 語あたりの難しさ」との順位相関を調べる．相関の指標にはスピアマンの順位相関係数[*7]を用いる．実験 1 における提案手法との比較対象として (1)[三好 10] の従来手法 (2) 評価表現度 (式 7 の第 1 項のみ) で推定した場合 (3) 評価表現度+tf 値のみを用いた専門度の場合 (4) 専門度 (式 7 の第 2 項のみ) で推定した場合を用いる．また，共起クラスタリング時の相互情報量の閾値 α を徐々に変化させたとき，そして難易度の計算に全てのレビューを用いた場合と，参考度を適用した場合で，提案手法の精度がどのように変化するのかを調べた．従来手法のパラメータは従来手法を参考にし，それぞれ p_a=0.4，p_d=0.8 とした．提案手法のパラメータ $Const$ は 0.001 と設定した．

∗5 評価表現辞書には「見やすい」という表現を入れているため
∗6 http://mediamarker.net/

∗7 スピアマンの順位相関係数 r_s は以下のように表せる．
$r_s = 1 - \frac{6}{N(N^2-1)} \sum_{i=0}^{N} (x_i - y_i)^2$

表 4 各手法と正解データ 1 とのスピアマン順位相関係数の比較

手法		MI>2.0	2.5	3.0	3.5	4.0	4.2	4.5	5.0	5.5
(1) 従来手法	0.633	-	-	-	-	-	-	-	-	-
(2) 評価表現度 (信憑性考慮)	0.767	-	-	-	-	-	-	-	-	-
(3) 評価表現度＋専門度 (tf)(信憑性考慮)	0.733	-	-	-	-	-	-	-	-	-
(4) 専門度 (信憑性考慮)	-	0.700	0.700	0.617	0.800	0.867	0.850	0.617	0.683	0.683
(5) **提案手法 (信憑性考慮)**	-	0.767	0.767	0.767	0.850	0.800	0.800	0.833	0.733	0.733
(6) 評価表現度 (信憑性考慮せず)	0.800	-	-	-	-	-	-	-	-	-
(7) 専門度 (信憑性考慮せず)	-	0.600	0.600	0.433	0.650	0.417	0.417	0.533	0.517	0.517
(8) 提案手法 (信憑性考慮せず)	-	0.800	0.800	0.800	0.833	0.717	0.717	0.667	0.717	0.717

5. 実験結果・考察

5·1 提案手法と他手法の精度比較

表 4 に実験 1 の結果を示す．ここでは，MI の閾値 α を変化させた場合の順位相関の値を，各手法と利用したレビューの信憑性毎に示している．すべてのパターンにおいて，提案手法は従来手法より高い相関係数を得ている．また，提案手法の第 1 項のみを難易度付与に用いた場合，第 2 項のみを難易度付与に用いた場合と比較しても提案手法が高い傾向にあることが分かる．スピアマンの順位相関はデータの半数の順序が入れ替わると相関係数がほぼ 0 となる評価指標であるが [加藤 03]，高い正の相関が得られたことから本手法は有効な推定手法であると言える．

次に，Spearman の検定表を用いて，実験結果から得られた相関係数が有意に 0 ではない (偶然起こったものではない) ことを確かめる．帰無仮説 H_0 を「提案手法と正解データの相関係数は 0 である」とする．検定表から提案手法の全てにおいて p <0.05[*8]で帰無仮説を棄却でき「提案手法と正解データ間には相関性がある」と言える．また，KS 検定（片側）により専門度の項だけで推定した場合の相関係数と提案手法で推定した場合の相関係数との間に有意水準 5% で中央値に差があることが認められた[*9]．このことから提案手法はより正しく難易度推定ができていると言える．

5·2 難易度と評価表現・専門用語の関係

表 4 から評価表現度は正解データとの相関係数が 0.767 と高いため，難易度が低い本ほど評価辞書内の表現が多く，難易度が高い本ほど少ない傾向にあると言える．

表 5 は本の難易度を正解データ 1 に沿って左から高い順に並び替え，各本の 1 文あたりに出現する専門用語の頻度を表している．この表から，正解データ 1 との順序相関係数 0.717 が得られ，難易度が高い本ほど専門用語が含まれている傾向にあることが分かる．

以上のことから，評価表現および専門用語の出現頻度は，本の難易度を推定するための指標として有効であると考えられる．

5·3 共起クラスタリングの効果

表 4 の手法 (3) の精度は提案手法よりも低い結果となった．そこで，実験 2 で各本の難易度と，専門用語の難しさ/語との相関を取ったところ 0.433 と中程度の相関係数が得られた．この結果から，難易度が高い本ほど意味を捉えにくい用語が入っている傾向にあるとは言いにくく，全ての用語の tf 値のみを用いて難易度の計算をするとさらにノイズが入る危険性があることが分かった．図 6 は相互情報量の閾値 α を変化させたときの各本の専門用語 1 語あたりのの難しさを示している．なお，$\alpha = 9.5$ のときに $|C| = 0$ となり，$\alpha \geqq 9.5$ においては共起クラスタリングが行われないことがわかった．図 6 より，共起に基づき tf 値を平滑化させたことで，tf 値のみを使ったよりも MI が 3.5〜4.2 の閾値によるクラスタリング後の方が，式 7 第 2 項の順位相関が上昇している．それに伴い提案手法の精度も向上している．また，相互情報量の閾値を小さくしていくにつれ，あらゆる専門用語 w が互いに共起するようになり，ほとんどの専門用語において $Diff_{tech}(w)$ が同値になっていく．その結果，図 6 では MI<3.0 のときにレビュー内の専門用語の難しさ/語が各本ともほぼ等しくなっている．つまり，tf 値を平滑化させすぎると専門用語の難しさの情報がなくなり，手法 (2) との順序相関 0.767 と一致する．対して，MI> 5.0 のように閾値が高すぎると，共起クラスタリングされず，手法 (3) との順序相関 0.733 と一致していることが分かる．tf 値の傾向を崩さず平滑化する方法として対数を取ることも考えられるが，底を 2 とし式 7 に適用したところ，正解データ 1 と式 7 第 2 項の順位相関は 0.633 となった．よって，共起性による値の平滑化がより本の難易度との関連を持っている．

次に，共起クラスタリング後で tf 値がどのように変化したのかを確かめた．tf 値の用語の難しさと MI 適用後の難しさを，難しさ別に 15 個のクラスタに分け[*10]，クラスタの難しい順に並び替える．ndpm 法[*11][横森 04] により，tf 値と MI 適用後の用語の難しさの大小関係の類似度を求める．

$$ndpm(\succ_{tf}, \succ_{MI}) = \frac{error}{{}_{161}C_2} \tag{8}$$

*8 相関関係が誤差である確率が 5% 未満である

*9 平均値に着目しなかった理由は提案手法における相関係数の数値に正規性がなかったためである．

*10 共起クラスタリング後の専門用語の難しさは同値である用語が多いのでクラスタに分けた

*11 用語 $w1$,$w2$ があったときに $C(w1) \succ_{tf} C(w2)$ かつ $C(w2) \succ_{MI} C(w1)$ を error とする

表 5 本の難易度と専門用語の出現頻度との関係

本の種類	(F)	(I)	(E)	(A)	(G)	(B)	(D)	(C)	(H)
専門用語の個数/レビューの文数	1.311	1.257	0.731	1.092	0.671	0.667	0.610	0.606	0.969

tf 値と相互情報量適用後の専門用語の難しさの距離が変化しなければ，専門用語の難しさを崩さない (考慮した) 状態で値を変化させたことになる．結果は MI<3.5 は 0.265,MI<4.0 では 0.237，MI<4.2 では 0.235 と約 75% 大小関係に変化がなかったため，tf 値の傾向が崩れていないことが分かる．

今回，全ての専門用語の難しさを難易度計算の指標としたが，実験 2 で被験者がつけたスコアが 2 未満の比較的意味を捉えにくい専門用語のみの総和を調べたところ，本の難易度の正解データ 1 との相関係数が 0.433 から 0.772 となり，高い相関が得られた．このことから専門用語の難しさを同定し，用語の難しさ top N (N:自然数) だけを難易度の計算に使えば，精度が上がると考えられる．

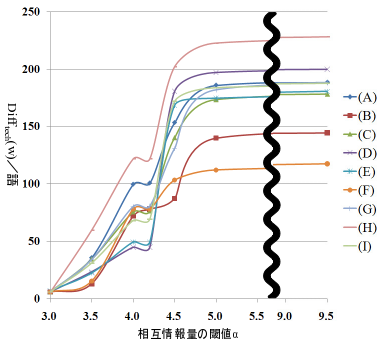


図 6 閾値の変化と専門用語の難しさ/語の関係

5·4 改訂版との比較

表 6 は実験 3 における相互情報量の閾値が 3.5 のときの，本 (J)・(K) の改訂前後と正解データ 3 の難易度順位を示している．また，() 内の数値は提案手法によって計算された難易度 $Diff(i)$ を表している．提案手法による本 11 冊の難易度と正解データ 3 との順位相関係数は 0.850 から 0.785 に変化した．低下の原因として表 6 のように本 (K) の改訂後と正解データの順位差が改訂前との順位差よりも大きくなったためである．また，正解データ 3 において改訂前後の難易度を同値としているので提案手法による難易度計算でも難易度が完全に一致しない限り，順位相関は低下すると考えられる．本 (J) では $Diff(i)$ が 0.636 と圧倒的に低く，一番易しい本と推定されており，改訂前後で難易度の開きが生じていることが分かる．これは 1 レビューあたりの行数が少ないレビューが多いことや評価表現が 1 文中に複数出現し，$Easy(i)$ の項が大きくなりすぎたためである．

今回はパラメータ $Const$ の値を経験的に定めたが，実験結果において $Const$ を大きくしていく ($Tech(i)$ の重みを増やす) ことで提案手法は $Easy(i)$ の異常値に対応できる．現状として 2 つの指標の信頼性を考慮しておらず，信頼性の高い指標に重みを強くすることで精度の向上が見込める．実験結果より $Tech(i)$ は相互情報量の閾値 α が変化しても順位相関が高い傾向にあり信頼性の高い指標と考えられるが断定はできない．よって，指標の信頼性を考慮したパラメータの決定方法は今後の課題となる．提案手法では評価表現の個数を数えているが，評価表現が含まれている行数を数えれば，評価表現が複数個出現する影響を抑えることができる．また，目次は改訂前後でも変化しないと予想できるので目次と組み合わせた難易度推定も考えられる．

表 6 本 (J)・(K) の改訂前後と正解データの難易度順位

	改訂前	改訂後	正解データ
本 (J)	7(0.8560)	11(0.6360)	9
本 (K)	6(0.8760)	4(0.9203)	7

5·5 信憑性の低いレビューを除いた効果

表 4 の手法 (5) と手法 (8) を比べると全てのレビューを難易度計算に用いると提案手法の精度が落ちることが分かる．また，提案手法の第 2 項である専門度を調べると，手法 (4) と比べて，手法 (7) のほうが精度が悪くなっていることも分かる．この原因を明らかにするために，専門用語の出現頻度と正解データ 1 との順序相関において，全てのレビューを用いた場合と信憑性が低いレビューを除いた場合を比べたところ，0.717 から 0.633 とわずかな低下が見られた．また，レビューの信憑性を考慮しない場合，共起クラスタリング後の専門用語の難しさ/語と正解データ 1 との相関が，悪い方向に変化したことも分かった．特に，本 (F)，(H) は信憑性の低いレビューの割合が多く，専門用語の難しさ/語が正解データ 1 との相関が悪くなるように変動したため，これら 2 冊のレビューを調べた．調査方法として信憑性の低いレビュー (低レビューと書く) とそうでないレビュー (高レビューと書く) に分け，正解データ 2 に従ってそれぞれのレビュー内に出現する専門用語の難しさの平均をとった．結果は，本 (F) では高レビューの方が低レビューより専門用語が難しくなり，本 (H) では高レビューの方が低レビューより専門用語が易しくなった．また，専門用語の難しさを正解データ 2 の代わりに相互情報量の閾値 $\alpha > 4.2$ での「共起クラスタリング後の専門用語の難しさ」を適用しても同様の結果となった．この理由は以下のように考えられる．本 (H) は正解データ 1 によると最も難易度の低い本である．この本の低レビューには専門用語「機械語」や「共用体」など正解データ 2 が難しいと判断した用語が

含まれていた．そのため，低レビューの方が専門用語が難しくなっていると考えられる．難易度が最も易しい本にも関わらず，難しい専門用語が含まれているレビューはユーザーにとって支持されていなかった．一方，本 (F) は最も難易度の高い本とされている．高レビュー内には，「リンカー」,「プリプロセッサ」など正解データ 2 が難しいと判断した用語が含まれていた．一方，低レビューは「C 言語」や「コード」といった正解データ 2 が易しいと判断した専門用語しか含まれていなかった．これらのように，低品質なレビュー内の専門用語の難しさは本の難易度に反比例する傾向にあり，それを除くことにより専門用語の難しさが正解データ 1 との相関をより持つようになったと思われる．この問題は 2 冊以外の本にも影響していると考えられるが，低レビューの割合が低いことやレビュー内に含まれる専門用語の個数が多いため，影響を抑えることが出来たと考えられる．よって，参考度適用の効果はレビューが比較的多い本には見られなかったが，低レビューの割合が高く，かつレビューが少ない本にはあったと言える．

5·6 提案手法による 22 冊の本の難易度ランキング

本論文では 11 冊の本に対して難易度を付与し，考察を行った．我々はさらに本の冊数を増やした時にどのようなランク付けになるのかを調べた．表 7 は本の冊数を 22 冊にしたときの難易度のランキングを表している．なお，実験で用いた本 11 冊には記号を付加している．7 位の『独習 C 第 4 版』では 8 件のレビューがあり，そのうち信憑性が低いレビューの排除によって 3 件のレビューしか得られなかった．レビュー数が極端に少ない場合は難易度の計算をしても信頼性は低い．提案手法では難易度計算として使うレビューを参考度が 0.4 以上で投票数が 5 以上としたが，レビューの新しさは考慮されていない．参考度と投票数によって排除されたレビューの中には「4 人中 4 人」のような参考度は高いが投票数が少ないレビューが見られた．このようなレビューは最近書き込まれたものが多く，信憑性が高い可能性のあるレビューとみなすことができる．よって，参考度が高く投票数が少ないレビューの中で，書き込まれた日付が最近のものを排除しないことで，この問題を解決できると考えられる．

本の難易度は「レビュー」や「商品の説明」を参照するだけでもある程度の情報は汲み取れる．本論文では表 7 のように難易度情報をランキングで可視化したという点で「レビュー」よりもユーザーにとって直感的でわかりやすくなる．Amazon の検索結果では膨大な数のレビューを保持している本が見られ，それらのレビューを読み終えるまでには時間がかかってしまう．また，中には信憑性の低いレビューが含まれている危険性があり，全ての情報を鵜呑みにはできない．さらに，商品の説明では絶対的難易度が記述されているが，ある別の本より難しいのかといった情報はない．Amazon などでは「おすすめサービス」や「検索結果一覧」により多数の本をユーザーに提示しているので，ユーザーはどうしても他書と比較してしまう．よって，本論文で求められた難易度ランキングは，学術本を比較し，選択を支援する指標として大変有用であると考える．

表 7 22 冊の C 言語に関する本の難易度ランキング

順位	本のタイトル	難易度
1	エキスパート C プログラミング (F)	0.9660
2	C 言語プログラミング (Computer Science Textbook)	0.9597
3	C 言語ポインタ完全制覇 (I)	0.9460
4	プログラミング言語 C 第 2 版 ANSI 規格準拠 (A)	0.9442
5	C プログラミング専門課程	0.9399
6	C 言語 (入門ソフトウェアシリーズ)	0.9234
7	独習 C 第 4 版	0.9230
8	やさしい C++まずは「C 言語」からはじめよう !!	0.9215
9	猫でもわかる C 言語プログラミング第 2 版 (K)	0.9203
10	独習 C 第 3 版	0.9167
11	新・C 言語入門シニア編 (E)	0.9114
12	10 日間でおぼえる C 言語入門教室	0.8770
13	猫でもわかる C 言語プログラミング (D)	0.8760
14	これならわかる C 入門の入門	0.8710
15	定本明解 C 言語 (第 1 巻) 入門編	0.8650
16	やさしい C 第 2 版 (C)	0.8560
17	新版 明解 C 言語 入門編 (G)	0.8380
18	新・C 言語入門ビギナー編 (B)	0.8260
19	新・C 言語入門スーパービギナー編 (H)	0.8130
20	C 言語 (1) はじめての C プログラミング	0.7520
21	C の絵本-C 言語が好きになる 9 つの扉	0.7410
22	やさしい C 第 3 版 (J)	0.6360

6. お わ り に

本論文では本を検索・推薦する属性値として嗜好情報だけでなく本の難易度を考慮する必要があると考え，本の難易度を推定する手法を提案した．レビュー中に出現する評価表現と専門用語を用いて推定を行い，実験により従来手法よりも高い精度が得られた．レビューは 1ヶ月，半年，1 年後で増えていくため，今後は本の難易度がどのように変化していくかを調べていく予定である．

既存のシステムでは検索結果の並び替えに難易度というキーがなかったが，提案手法によって推定された難易度にしたがって並び替えることで，初学者や次の本を探すユーザーにとって有用であると考えられる．さらに検索結果だけでなく推薦方法にも違いが生まれる．Amazon のおすすめサービスでは嗜好情報に基づき，関連のある本を推薦しているが，本研究では今閲覧している本の $\pm\alpha$ の難易度である本を推薦できるため，学術本の選択支援に有用である．

また，難易度を数値化することで，難易度の高い本を読んでいる人は知識量があるなど，ユーザーの知識量推定が可能になる．すると知識量を属性値とした協調フィ

ルタリングを行うことで知識量の類似したユーザーが理解できた，もしくは少し難しいと評価した本を推薦することができる．また，知識量が豊富であるユーザーが分かることにより，そのユーザーにアドバイスをもらうといった学習支援につながることも期待できる．本論文ではジャンルを C 言語に絞ったが，他分野の学術本のレビューにも評価表現・専門用語が含まれる．また，同じように tf 値に大差があると思われるため，他ジャンルの学術本にも有効であると考えられる．

◇ 付　　録 ◇

A. 専門用語 161 語一覧

16 進数 (2), 2 進数 (2), ANSI(13), BASIC(4), C(230), C++(39), C コンパイラ (6), C 言語 (381), DOS(4), FILE(3), Fortran(2), OS(4), UNIX(6), c(6), char(7), const(5), c 言語 (5), enum(11), error(2), for(2), for 文 (3), if(3), include(2), int(8), main(3), printf(6), printf 関数 (3), return(5), scanf 関数 (2), static(3), stdio(2), stdout(2), TRUE(2), type(2), unix(2), void(2), ＡＮＳＩＣ (2), C (262), ＤＯＳ (3), ＵＮＩＸ (3), アクセス (5), アスキー (2), アセンブラ (6), アドレス (12), アルゴリズム (18), インクリメント (3), エディタ (3), エラー (12), オブジェクト (3), オブジェクト指向 (7), オプション (5), キーワード (4), キャスト (6), キャラクター (2), コーディング (24), コード (79), コメント (12), コンパイラ (30), コンパイル (23), コンパイルエラー (5), コンピュータ (14), コンピューター (4), システム (5), スコープ (2), スタック (2), ソース (50), ソースファイル (3), ソフト (15), ソフトウェア (7), タイプ (2), テキスト (18), テキストモード (4), データ (12), データ型 (2), デバッグ (2), バイト (2), バイナリモード (2), バグ (10), ビット (2), ビット演算 (2), ファイル (17), ファイル入出力 (7), フローチャート (4), プリプロセッサ (5), プログラマ (35), プログラマー (11), プログラミング (208), プログラミング言語 (26), プログラム (238), プログラム言語 (5), プロンプト (2), ヘッダー (3), ヘッダファイル (4), ポインタ (245), ポインター (2), マクロ (6), マニュアル (7), メモリ (32), ユーザー (4), ライブラリ (14), ライブラリ関数 (4), ラベル (2), リスト (6)、リッチー (4), リンカー (2), ループ (3), ワード (2), 依存 (3), 移植性 (4), 引数 (10), 演算 (9), 演算子 (4), 関数 (66), 機械語 (2) 共用体 (3), 繰り返し文 (4), 型 (2), 計算機 (4), 言語 (251), 互換性 (2), 語 (4), 構造化プログラミング (2), 構造体 (40), 構文 (4), 行 (5), 項 (5), 指針 (2), 式 (2), 実行 (37), 手順 (5), 出力 (6), 処理系 (2), 初期化 (2), 数 (26), 数字 (2), 制御構造 (2), 制御文 (5), 宣言 (9), 代入 (8), 値 (7), 定義 (6), 定数 (2), 底 (2), 動的 (2), 日本語 (23), 入力 (9), 配列 (40), 配列名 (3), 否定 (4), 標準ライブラリ関数 (2), 評価 (25), 文字 (19), 文字列 (9), 文法 (57), 文脈 (2), 変数 (20), 変数名 (2), 翻訳 (12), 名前 (5), 流れ (11), 例外 (2)

◇ 参　考　文　献 ◇

[M.Kleinberg 98] M.Kleinberg, J.: Authoriative Sources in a Hyperlinked Environment, *Proceedings of the ACM-SLAM symposium on Discrete Algorithms* (1998)

[Nakatani 09] Nakatani, M., Jatowt, A., and Tanaka, K.: Easiest-First Search:Towards comprehension-based Web Search, *In Proceedings of the 18th ACM Conference on Information and Knowledge Management(CIKM 2009)*, pp. 2057–2060 (2009)

[Nishihara 05] Nishihara, Y., Sunayama, W., and Yachida, M.: Information Acquiring Support System based on KeyWords' Continutiy and Informational Difficulty, *in Proc. of International Conference on Human-Computer Interaction(HCI2005), Las Vegas* (2005)

[Sato 08] Sato, S., Matsuyoshi, S., and Kondo, Y.: Automatic assessment of japanese text readability based on a textbook corpus, *In 6th LREC* (2008)

[伊藤 07] 伊藤 雅弘, 中山 浩太郎, 原 隆浩, 西尾 章治郎：Wikipedia のリンク共起性解析によるシソーラス辞書構築, 情報処理学会論文誌. データベース, Vol. 48, No. 20, pp. 39–49 (2007)

[横森 04] 横森 励士, 梅森 文彰, 西 秀雄, 山本 哲男, 松下 誠, 楠本 真二, 井上 克郎：Java ソフトウェア部品検索システム SPARS-J, 電子情報通信学会論文誌.D-I 情報・システム, I-情報処理, Vol. 87, No. 12, pp. 1060–1068 (2004)

[加藤 03] 加藤 千恵子, 石村 貞夫：相関係数と回帰直線, 東京図書 (2003)

[三好 10] 三好 康夫, 入野 美弥：学術書籍の難易度を読者ネットワークから推定する試み, 電子情報通信学会技術研究報告. ET, 教育工学, Vol. 110, No. 67, pp. 19–24 (2010)

[小倉 08] 小倉 達矢, 宍戸 開, 今藤 紀子 山口 実靖, 淺谷 耕一：レビューサイトにおける良質なレビューの特性とそれを考慮した評判情報の抽出に関する一考察, 第 19 回データ工学ワークショップ (DEWS 2008)B8-5 (2008)

[松尾 02] 松尾 豊, 石塚 満：語の共起の統計情報に基づくキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol. 17, pp. 217–223 (2002)

[杉木 08] 杉木 健二, 松原 茂樹：消費者の意見に基づく商品検索, 情報処理学会論文誌, Vol. 49, No. 7, pp. 2598–2603 (2008)

[倉島 07] 倉島 健, 別所 克人, 戸田 浩之, 内山 俊郎, 片岡 良治, 奥 雅博：比較評価情報に基づくランキング手法, 日本データベース学会論文誌 (DBSJ Letters), Vol. 6, No. 1, pp. 5–8 (2007)

[相澤 00] 相澤 彰子, 影浦 峡：著者キーワード中での共起に基づく専門用語間の関連度計算法, 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol. 83, No. 11, pp. 1154–1162 (2000)

[中川 03] 中川 裕志, 湯本 紘彰, 森 辰則：出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理 = Journal of natural language processing, Vol. 10, No. 1, pp. 27–45 (2003)

[中谷 08] 中谷 知博, 星野 准一：経験的価値の分類に基づくゲーム推薦システム, 情報処理学会研究報告. EC, エンタテインメントコンピューティング, Vol. 2008, No. 129, pp. 49–56 (2008)

[中條 04] 中條 清美, 白井 篤義, 内山 将夫：日英パラレルコーパスを構成するテキストの難易度分類に関する研究, 日本大学生産工学部研究報告.B, 文系 37, pp. 57–68 (2004)

[藤村 05] 藤村 滋, 豊田 正史, 喜連川 優：文の構造を考慮した評判抽出手法, 電子情報通信学会第 16 回データ工学ワークショップ (DEWS 2005) (2005)

[平林 97] 平林 雅英：新 ANSI C 言語辞典, 技術評論社 (1997)

[鈴木 09] 鈴木 健太, 濱川 礼：5N-1 他人のコンテンツ評価を用いたユーザの嗜好推測に基づくコンテンツ推薦 (推薦, 学生セッション, データベースとメディア), 全国大会講演論文集, Vol. 71, No. 1, pp. 575–576 (2009)

〔担当委員：平嶋　宗〕

2011 年 8 月 4 日　受理

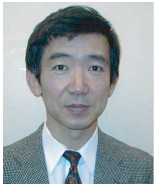
著　者　紹　介

中山　祐輝

2010 年金沢大学工学部情報システム工学科卒業．2012 年 3 月，同大学大学院自然科学研究科修士課程修了．2012 年 4 月より東京工業大学大学院博士課程に在籍．自然言語処理，情報検索・推薦，評判分析，Web マイニング等に興味がある．言語処理学会会員．

南保　英孝

1999 年金沢大学大学院自然科学研究科博士課程修了．博士（工学）．同年同大学工学部電気・情報工学科助手．2002 年同学部情報システム工学科講師．現在，金沢大学理工学域電子情報学類講師．センサ情報処理とそのアプリケーションに関する研究を行っている．電気学会，電子情報通信学会，情報処理学会会員．



木村　春彦（正会員）

1979 年 東北大学大学院工学研究科博士課程修了．博士 (工学)．同年，富士通（株）入社．1980 年 金沢女子短期大学講師．1984 年 金沢大学経済学部助教授．現在，同大学大学院自然科学研究科教授．この間，最適コード変換やプロダクションシステムの高速化に関する研究に従事．電気学会，電子情報通信学会，情報処理学会各会員