ISSN 2186-5647

P-25

レビューを用いた学術書の難易度の推定

日大生産工（学部）　○山本　修平
日本大学生産工　関　亜紀子

1 まえがき

近年、ECサイトでは商品概要や購入履歴などをもとに、自分に合った商品や関連商品などのお勧めを知ることが出来る。こうした推薦手法は書籍販売でも活用されているが、専門度や難易度が異なる学術書の場合は、難易度に応じてソートできるなど、読者の知識レベルに応じて選書できる機能が必要とされている[1]。

これに対して三好ら[2]は、利用者の嗜好だけではなく、読書履歴をもとに、読者間のネットワークを構築し、協調フィルタリングによりユーザの習熟度とコンテンツの難易度の推定を試みている。また、舟木ら[3]は、目次および書名を用いてコンピューター関連書籍の難易度および類似度の分類を行っている。中山ら[1]は、レビューに含まれる評価表現と専門用語を用いた書籍の難易度推定手法を提案している。

我々は、ECサイトの利用者が書籍を選ぶ際、タイトルだけではなく本の概要文、読者レビューを見て読む本を選ぶという点に着目する。従来の方法では限られた分野のみでしか分類をすることができなかった。そこで、ユーザの用途によって基本辞書を変えることで、様々な分野の書籍での分類を可能にする方法について検討する。

2 学術書のレビューの傾向

学術書レビューには、図1の(a)と(b)に示すように、実際に使用した読者の使用感を示すキーワードが随所に表れている。下線部のキーワードなどに着目すると、初心者向けで基本的な解説が多く含まれる本なのか、実用的な記述が多く含まれる本なのかをレビューから読み取ることができる。そこで、同一のキーワードでヒットする書籍を対象に、書籍の難易度に応じてレビューに出現する専門用語の傾向に特徴がみられるか、termmi[4]を使って比較した。

図2は、15冊を対象にレビュー内の専門用語の類似度を比較した結果である。ここでは、「C言語」もしくは「C++」で検索される書籍の中

・初心者の学生さんがつまずかないように丁寧に構成してあり、教科書として利用しやすい．絵による説明も適切である．

(a) 初心者向けの本のレビュー例

・この本で、IDEを使わないでどう効率よく開発をすればいいのか少し実感できました。
・あらゆる言語で便利な開発環境やテストスートが出来ており、C言語とて例外ではありません。

(b) 実用的な本のレビュー例

図1. 使用感を示すレビュー表現

	A1との類似度	C1との類似度	書籍のタイトル
A1	1	0.180	新板C言語 プログラミングレッスン
A2	0.225	0.119	やさしいC　第3版
A3	0.120	0.112	1日で解るC言語
A4	0.048	0.033	新・C言語のススメ
A5	0.322	0.114	新・明解C言語 入門編
B1	0.085	0.067	C言語によるアルゴリズムとデータ構造
B2	0.161	0.163	Cプログラミングの基礎
B3	0.101	0.220	すぐわかるC/C++
B4	0.154	0.180	新・明解C言語 実践編
B5	0.129	0.292	C実践プログマミング 第3版
C1	0.180	1	開発ツールを使って学ぶ！C言語プログラミング
C2	0.137	0.306	プログラミング言語C++ 第4版
C3	0.153	0.350	モダンC言語プログラミング
C4	0.204	0.270	C言語によるスーパーLinuxプログラミング
C5	0.079	0.177	明快入門コンパイラ・インタプリタ開発

図2. レビュー間の類似度

から、入門書や基礎的なことが記載されている書籍をグループA、中級者向けの本、実践的な本をグループB、更に実践的で専門的なことが記述されている本をグループCとした。事前に、各グループに対応する書籍を5冊ずつ選定し、それぞれ最大5件ずつレビューを収集した。

図2より、Aグループに属している書籍A1との類似度の高い書籍がA5、A2であることがわかる。また、書籍C1との類似度の高い書籍がC3,C2であることがわかる。このことから、同一のキーワードで検索可能な書籍のレビューは、その難易度に応じてレビューに現れる表現が類似する傾向にあり、書籍の難易度の分類に活用できると考えられる。なお、書籍A4はレ

Difficulty Determination Method for Academic Books using Reviews
Shuhei YAMAMOTO and Akiko SEKI

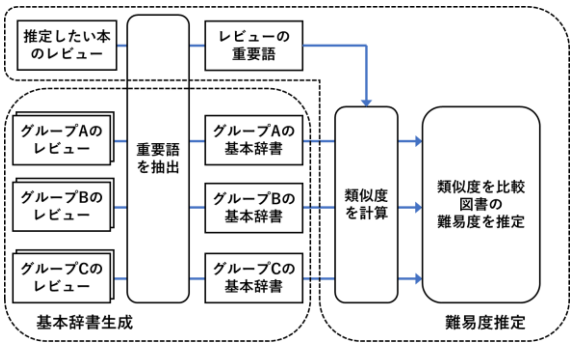

図 3. 難易度推定の処理概要

グループA		グループB		グループC	
本	41.00	C	85.05	C	132.90
言語	38.26	本	38.26	言語	64.23
C言語	20.50	言語	30.06	本	56.03
プログラミング	17.76	内容	23.23	開発	47.84
C言語プログラミング	14.79	C	22.19	環境	39.44
C言語	13.67	C言語	21.86	C言語	38.26
説明	13.67	プログラム	20.50	方法	29.58
テキスト	10.63	版	19.72	コンパイラ	27.12
必要最低限	9.86	アルゴリズム	19.72	ツール	24.65
初心者	9.57	コード	14.18	本書	24.60
内容	8.20	実践	14.18	書籍	24.60
入門書	8.20	本書	13.67	プログラミング	21.86
入門編	7.09	関数	12.40	コード	21.26
基本	7.09	初心者	10.93	知識	21.26
星	7.09	学習	10.93	内容	20.50
最初	7.09	文法	10.93	ライブラリ	19.49
他	6.83	基本	10.63	人	17.76
ポインタ	5.32	実践編	10.63	プログラム	15.03
反面	4.93	データ構造	9.86	仕様	14.79
書き込み用	4.93	プログラミング	9.57	技術	14.79

図4. 各グループの基本辞書の一例

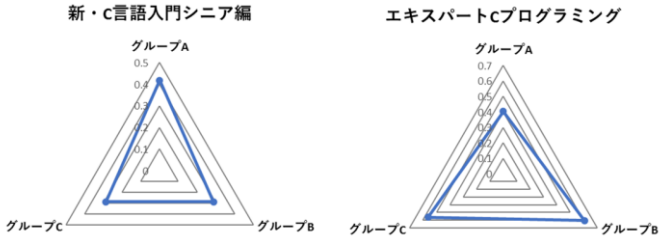

図5. 難易度推定結果の例

ビュー件数が少ないために、十分な精度に至らなかった。

３．難易度推定の流れ

提案する書籍の難易度推定の処理の概要を図3に示す。まず、ユーザがECサイトなどから選定した本を基礎的な本、実践的な本、専門的な本などの３つのグループに分類し、これらのレビューをシステムへの入力とする。基本辞書の生成では、各グループに属する書籍のレビュー全体から専門用語を取り出し、TF-IDF法を用いて各用語の重要度を計算する。その後、重要度が4.0以上の用語をグループを特徴づけるキーワードとして抽出し、分類基準となる基本辞書とする。難易度推定では、難易度を推定したい書籍に対して投稿された全レビューから専門用語のみを抽出し、それらの重要度をTF-IDF法を用いて計算する。その後、抽出した専門用語と各グループの基本辞書内の専門用語との類似度をベクトル空間法を用いて計算し、最も数値が高かった辞書の難易度を対象の書籍の難易度とする。

４　まとめと今後の課題

本報告では、学術書のレビューに現れる専門用語の特徴を活かした難易度推定手法の提案を行った。

提案手法に基づいて作成した基本辞書の例と、それらを用いた書籍の難易度推定の結果を図4と図5に示す。図4は、図2に示す書籍を対象に作成した基本辞書内の専門用語の一例である。グループAの辞書とグループCの辞書を比較すると、重要度が大きく離れている。レビューの合計文字数は、グループAが 3687文字だったのに対し、グループCは10009文字だった。このことから文章量の差で重要度の差が生まれると考えられる。

図5は、2冊の学術書を対象に難易度推定を実施した結果である。「新・C言語入門シニア編」は、グループBの難易度を想定していたが、推定結果はグループAと判断された。また、「エキスパートCプログラミング」は、グループCの難易度を想定していたが、グループBと判断された。これは基本辞書作成に使った書籍の分類分けが正しく行われていなかったことが原因として考えられる。

今後は、これらの問題点を改善すると共に、目次や概要などを用いた場合と比較を行い、推定結果の妥当性を検証する。

「参考文献」

1) 中山 祐輝, 南保 英孝, 木村 春彦, レビュー情報を用いた学術本の難易度推定, 人工知能学会論文誌, Vol. 27 (2012) No. 3, p.213-222
2) 三好 康夫, 入野 美弥, IRINO Miya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-110 (67), (2010) p. 19-24
3) 舟木 類佳, 黒田 久泰, 難易度及び類似度を用いたコンピューター関連書籍推薦システムの開発, 情報処理学会研究報告, 2014-NL-215 (8), (2014) p. 1-6
4) Windows用テキストマイニングツール “termmi”, http://gensen.dl.itc.u-tokyo.ac.jp/termmi.html (2017)