

レビューを利用した読者の印象に基づく書籍推薦手法

萱田真史[†] 村井聡一[‡] 牛尼剛聡[‡]

[†]九州大学〒815-8540 福岡県福岡市南区塩原 4-9-1

[‡]九州大学大学院芸術工学研究院〒815-8540 福岡県福岡市南区塩原 4-9-1

E-mail: [†] hoooomin@hotmail.com, [‡] ushiama@design.kyushu-u.ac.jp

あらまし Amazonに代表されるオンライン書店では、協調フィルタリングを用いた、書籍の推薦が行われている。しかし、単純な強調フィルタリングを用いた推薦手法では、同作家や同ジャンル、人気作品などの意外性のない対象が推薦される場合が多い。そこで、本研究では、書籍のレビューを用いて書籍への印象に関する単語を抽出し、書籍間の印象に基づいた類似性を利用することで、意外性のある書籍の推薦を実現することを目指す。

キーワード 書籍, 推薦, レビュー

1. はじめに

近年、iPad や kindle などの電子書籍が閲覧できるデバイスの普及とともに、ユーザに提供される書籍数が飛躍的に増加している。一方で、大量の書籍の中からユーザが興味を持てる書籍を探すことが困難になっている。そのような背景の下、効果的な推薦を行う手法の重要度が増加している。Amazon などのオンライン書店では、協調フィルタリングを用いた推薦手法が一般的に行われている。

単純な協調フィルタリングによる書籍推薦にはいくつかの問題点が存在する。その一つとして、同著者による書籍や同ジャンルの書籍、または一般敵に人気度の高い書籍の推薦が圧倒的に多く、意外性に欠けることがある。例えば Amazon における村上春樹の「ノルウェイの森 (上)」の書籍の紹介ページでは、「この商品を買った人はこんな商品も買っています」という項目には、ユーザベースの協調フィルタリングによる推薦手法により、「海辺のカフカ(上)」、「海辺のカフカ(下)」、「ノルウェイの森(下)」等の書籍が推薦される。これらの書籍は総て同一著者によって執筆された有名な作品であり、意外性が低い。

そこで、本研究では書籍の内容の類似性や同一著者の書籍等の推薦を避け、意外性のある推薦を行うために、書籍に対するユーザの印象を推測し、それを書籍間で比較し、推薦に用いる手法を提案する。

本稿は 6 章から構成される。2 章では関連研究について述べる。3 章では提案手法のアプローチについて述べる。4 章では提案手法の詳細について述べる。5 章では提案手法の実験と考察による評価を行う。6 章ではまとめと今後の展望を述べる。

2. 関連研究

これまでにも、意外性をもつ推薦を行うために、いくつかの手法が提案されてきている。

近藤ら[1]は「意外性のある情報は、情報全体からそのユーザにとって有益でない情報を除外することで、取得できる可能性が高くなる」と考え、有益でない情報を除いたものが意外性のあるものという仮定に基づいた推薦手法を提案している。

村上ら[2]は「利用者のアクセスの習慣から外れており、かつ利用者の嗜好に適合したコンテンツを推薦すると利用者は意外に感じるであろう」との仮定のもと、嗜好モデルと習慣モデルとの差異で意外性を推定している。

しかし、上記の推薦手法では、類似したユーザの嗜好を基にするなど、協調フィルタリングを利用しているのに対して、本研究では、協調フィルタリングを利用せず、書籍のレビューの特徴に基づいて、書籍がユーザに与える印象の類似性を求め、意外性のある推薦を行う。

3. アプローチ

本研究では意外な書籍の推薦に対するアプローチとして、書籍の内容自体の類似性をではなく、書籍のレビューの類似性を基に推薦を行うことを考える。本研究では、レビューからユーザが書籍に対して感じた印象を抽出し、その印象の類似性を比較することにより、書籍の内容の類似性に依存しない推薦を行う。

本研究で想定する推薦には、ユーザにお気に入りの書籍 1 冊を指定してもらい、その書籍を基に、その書籍と印象の類似した書籍を推薦するというものである。例えば、あるユーザのお気に入りの書籍 A に対して、書籍 A と内容や背景、著者などは異なるが似た世界観を持った書籍 B が存在するとする。それらの書籍間には、本文よりもレビューに類似性があると考えられる。なぜなら、書籍のレビューには読者が書籍に感じた印象に関する記述が多く含まれていると考えられるからである。このような考えに基づくと、一般的に行われ

ている内容の類似性からは得られない、ユーザの印象に対する類似性に基づいて書籍を選択でき、一般的に知られていない意外性のある推薦を実現できると考えられる。図 1 に書籍の内容の類似性と、書籍に基づいて得られるユーザの印象の類似性の概念図を示す。

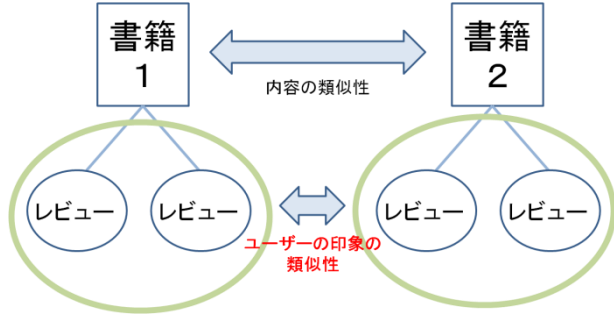


図 1：書籍の類似性と印象の類似性

4. 提案手法

本文の内容に依存しない推薦を行うために、内容に関わる単語をできるだけ省く必要がある。本研究では、書籍とその書籍に対するレビューに対して、形態素解析を適用し、それらを構成する単語集合を抽出する。そして、得られた単語集合のうち、形容詞のみを対象とする。これによって得られた、レビューに用いられている形容詞を、ユーザの書籍に対する印象を表わす語と考える。

以上の手法で抽出した各語の重みづけとして TF・IDF 法を用いる。

$$\text{tfidf} = \text{tf} \cdot \text{idf} \quad (1)$$

$$\text{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (2)$$

$$\text{idf}_i = \log \frac{|D|}{|\{d: d \ni t_i\}|} \quad (3)$$

ここで TF とは注目する書籍 1 冊の全レビューにおいて、それぞれの形容詞の出現回数 n_{ij} をその書籍内における全形容詞数 $\sum_k n_{kj}$ で割ったものであり、単語出現頻度を表わす。また、IDF は、比較に用いた全書籍数 D を、それぞれの形容詞が出現した文書数 $\{d: d \ni t_i\}$ で割ったものに対数をとったものであり、逆文書出現頻度を表わす。したがって、多くのレビューに出現する形容詞は TF・IDF 値が低くなり、特定のレビューにしか出現しない形容詞は TF・IDF 値が高くなる。IDF を求める際、対数関数に 1 を足すことで IDF の重みを小さくし

ている。本研究では、レビューに多く用いられる単語を使用したいため、IDF の重みを小さくして実験を行う。

以上で計算した各単語の TF・IDF 値を基に、コサイン相関値を用いて二つの書籍間の類似度を求める。

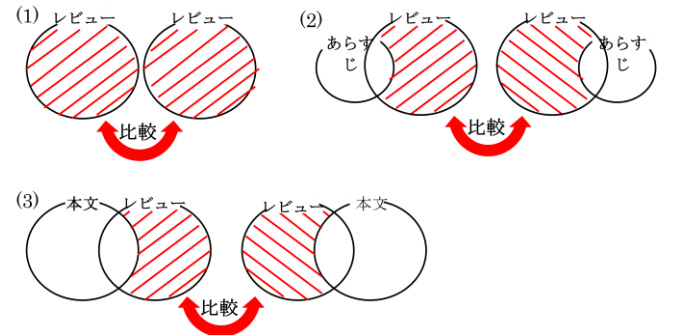
$$\begin{aligned} \text{sim}(D_i, D_j) \\ = \frac{w_{i1}w_{j1} + w_{i2}w_{j2} + \dots + w_{in}w_{jn}}{\sqrt{w_{i1}^2 + \dots + w_{in}^2} \times \sqrt{w_{j1}^2 + \dots + w_{jn}^2}} \end{aligned} \quad (4)$$

ここで、注目する書籍 D_i の TF・IDF ベクトルは、 $D_i = (w_{i1}, w_{i2}, \dots, w_{in})$ で表わされる。また、推薦候補となる書籍 D_j の TF・IDF ベクトルも同様に表わされる。

5. 評価・実験

提案手法の有効性を評価するために、以下の 3 つの手法で結果を比較する。

- (1) レビューをそのまま使用する。
 - (2) 書籍のあらすじにある語をレビューから除去する。
 - (3) 書籍の本文にある語をレビューから除去する。
- (2) では、レビューから抽出した形容詞のうち、あらすじにもある形容詞を除去し、残った形容詞で比較する。(3) では、書籍の本文は著作権があり、大抵の書籍において推薦に用いることができない。そこで、本文に無料で公開されている、青空文庫から著作権のない書籍を 40 冊選ぶ。それらの本文、レビューを用いて、本文にない形容詞をレビューから抽出し、出現頻度の高い 100 語を、レビューにおける印象を表わす語として使用する。



(1)、(2)については図 2 に示す書籍 10 冊で TF・IDF 値を計算した。書籍を選択する際、レビューを基に推薦することを考慮して、レビューの多い作品を実験に用いた。

表 1: 実験に用いた書籍

ID	タイトル	著者名
1	ノルウェイの森(上)	村上 春樹
2	1Q84 BOOK1	村上 春樹
3	重力ピエロ	伊坂幸太郎
4	ノルウェイの森(下)	村上 春樹
5	告白	湊 かなえ
6	夜は短し歩けよ乙女	森見 登美彦
7	オーデュボンの祈り	伊坂幸太郎
8	1Q84 BOOK2	村上 春樹
9	西の魔女が死んだ	梨木 香歩
10	死神の精度	伊坂幸太郎

表 2, 3 ではそれぞれの手法における小数点第 3 位を四捨五入したコサイン相関値を示した。各表の 1 から 10 までの数字は、書籍に付与した識別番号である。また、赤い文字は、コサイン相関値が高い上位 5 件を示している。

まず、表 2 の相関値上位 5 件を見ると、書籍 1、4 間、書籍 2、7 間、書籍 7、10 間の 3 書籍間において、同一著者であるという共通点が見られる。したがって、意外性のある推薦をするという目標に合致していない。

表を比較すると、書籍によって相関値が大きく変わるものとそうでないものの差が激しい。例えば、書籍 5 では相関値の変化が大きくとも 0.02 だが、書籍 3 では 0.04 以上の差異がある値が 5 か所ある。これは、あらすじの長さが書籍によって大きく差があり、レビューから除去される形容詞が大きく違うためであると思われる。あらすじを除去することによりコサイン相関値が最も大きく増加した書籍 3、7 間では表 4、5 に見られるように「重い」という単語以外での大きな差異は見られなかった。それにもかかわらず相関値に違いが出ているのは、書籍 3 のあらすじにおいて、レビュ

一の形容詞を多く除去したためだと考えられる。さらに、コサイン相関値が大きくなった書籍 3、7 間では、ともに伊坂幸太郎が著者であり、意外性のある推薦をするという目標にも合致しない。以上の結果より、あらすじに含まれる形容詞を除去することに有意性はないと考えられる。

6. おわりに

本論文では、ユーザのお気に入りの本 1 冊から次に読む書籍を探す際の書籍の推薦手法を提案した。同一著者や人気作品だけに留まらず、より読書の幅を広げ、ユーザの新たな興味を喚起するような推薦を行うことを目標としている。

提案手法では、初めに各書籍の全レビューから形容詞を形態素解析し、形容詞だけを抽出した。次に、抽出した形容詞でコサイン相関値を計算した。しかし、同一著者による書籍同士の値が大きくなったため、目標通りにはいかなかった。そして、より内容に関係のある単語を減らすため、あらすじを用いてフィルタリングを行った。しかし、書籍によってあらすじの量にばらつきがあり、これも意外性のある推薦を行うためには有意でないことが分かった。

今後の課題としては、手法(3)の実験・評価を行うことや、書籍によって抽出した形容詞が多様になるようにレビューの TF・IDF 値を出す際の IDF の重みを調整すること、著作権のない書籍からレビューに主に使われる形容詞集合の抽出方法の改善などを計画している。

参 考 文 献

- [1] Amazon : <http://www.amazon.co.jp/>
- [2] 近藤司, 伊藤真也, 原田史子, “意外性ある推薦のための非有益性に基づく情報フィルタリング”, FIT2010.
- [3] 村上知子, 森紘一郎, 折原良平, “推薦の意外性向上のための手法とその評価”, 人工知能学会論文誌 24 卷 5 号 G.

表 2: 手法(1)における書籍間の相関値

[illegible]

表 3：手法(2)における書籍間の相関値

	1	2	3	4	5	6	7	8	9	10
1		0.80	0.88	0.95	0.71	0.69	0.86	0.87	0.73	0.78
2			0.87	0.74	0.86	0.81	0.91	0.95	0.48	0.90
3				0.86	0.82	0.82	0.95	0.91	0.69	0.92
4					0.67	0.65	0.83	0.83	0.74	0.72
5						0.74	0.89	0.84	0.47	0.88
6							0.84	0.79	0.51	0.87
7								0.91	0.67	0.94
8									0.59	0.88
9										0.57
10										

表 4：手法(1)における書籍の TF・IDF 値上位

1	2	3	7
いい	面白い	面白い	面白い
面白い	多い	いい	すごい
すごい	いい	良い	いい
良い	おもしろい	重い	良い
多い	すごい	すごい	おもしろい
深い	早い	よい	よい
切ない	良い	おもしろい	多い
よい	長い	多い	楽しい
美しい	難しい	深い	悪い
暗い	よい	楽しい	凄い

表 5：手法(2)における書籍の TF・IDF 値上位 10 語

1	2	3	7
いい	面白い	面白い	面白い
面白い	多い	いい	すごい
すごい	いい	良い	いい
良い	おもしろい	すごい	良い
多い	すごい	よい	おもしろい
深い	早い	おもしろい	よい
切ない	良い	多い	多い
よい	長い	深い	楽しい
美しい	難しい	楽しい	悪い
おもしろい	よい	かっこいい	凄い

表 6：各書籍のあらすじ除去における単語除去数

書籍番号	1	2	3	4	5	6	7	8	9	10
レビューの 単語除去数	3	1	8	4	0	0	2	1	4	0