SIG-KSN-007-06
人工知能学会
第七回知識流通ネットワーク研究会

芥川賞受賞作品を利用した書籍売上予測の先行指標としてのBlog情報の活用手法

As a Leading Indicator of Book Sales Forecasting Method to Make Use of Blog Information Using Akutagawa Prize Book

菊田　剛　文 健哲　山田 隆志　吉川 厚　寺野 隆雄
Go Kikuta　Geun Chol Moon　Takashi Yamada
Atsushi Yoshikawa　Takao Terano

東京工業大学大学院総合理工学研究科

Abstract: This paper presents a Book sales tendency by analyzing the Akutagawa Prize Books from 2005 to 2009 in Japan. Why we chose akutagawa prize books is it has a very impressional effects for sales when announcing what book got this award. We analyzed the tendency of Akutagawa Prized Books text information, and the Cross-correlation between book sales and the number of references in Books using volume and blog text information. As a result, We found Akutagawa Prized Books sales and blog information has high cross-correlation and it may be useful when predicting book sales.

1　はじめに

出版業界全体の問題として, 書籍の返品率が非常に高い水準にある [1] という事が共通の課題として認識されている. 返品率が高い理由としては, 出版業界における制度の問題や, 適正な本の需要予測が出来ていない事などが理由として上げられるが, そのような問題を解決する上で適正な本の需要予測手法の確立を行う事は, 非常に重要である.

近年, Blog はインターネットメディアの中でも, 最も利用される媒体の一つとなっている [2]. Blog の利用手法としては, 様々な事象に対する自分の意見を述べる場として利用される事が多く, 人々の購買決定の要因としても強い影響を及ぼしている [3]. それらの情報を人々の集合知として活用する事により, 様々な商品に対する意見の指標として利用する事が可能だと考えられる. Blog を利用した研究には,Blog 情報を先行指標として利用して売上を予測する研究 [5, 7] が存在し, Gruhl ら [5] は Blog を利用することで, 売上の急激な伸びを予測することが可能だと示唆しており, また,Yang ら [7] は,Blog の感情情報を利用した ARSA モデルを提案しており, 従来の AR モデルより精度よく売上を予測可能な事を示唆している.

そのような背景を踏まえ, 我々は書籍の Blog 情報を先行指標とした書籍の需要予測モデルを作成する事を目指している.

2　関連研究

我々は過去の研究 [4] において, 国内書籍販売における先行指標としての Blog 情報の有用性を見出した. その結果,Blog 情報が書籍の売上に先行する際には映画や, ドラマといった外部イベントが強く影響する事を確認した. Blog 情報を利用した書籍の売上ランキング予測の研究は Gruhl ら [5] がある. この研究では Amazon のランキングデータと Blog 情報を組み合わせた需要予測の提案を行っている. 利用した Blog データは単純な Blog のヒット数であり, 予測効力はそれほど高くないものではないとの報告をしている. Wolfgang[6] らは本を対象としたオークションの価格予測モデルとして, AR モデルに本のカテゴリや送料を組み込んだモデルの提案を行っており, その結果, 本のカテゴリや送料は価格予測モデルの要素としてはそれほど有用でない事を示した. Yang[7] らは Blog 情報を利用して売上の予測をする手法として ARSA モデルを提案しており, そのモデルの評価を AR モデルなどと比較することで評価をしている. このモデルでは,Blog 情報を Probabilistic Latent Semantic Analysis(PLSA)[9] を利用した S-PLSA という手法で加工し, AR モデルと組み合わせた売上予測モデルの作成を行っており, 結果として,Blog の感情情報を利用した場合, そうでない場合よりも精度よく売上を予測できる事を示した. また,S-PLSA を改良した S-PLSA+[8] の提唱もしており, ARSA モデルに S-PLSA と比較した場合, より精度の高い結果を得られたことを

示した.

このような背景を踏まえ, 我々は Blog 情報を先行指標とした書籍の需要予測モデルを作成する事を目指している. Blog 情報を利用した書籍の需要予測モデルを作成するにあたって, まず Blog の先行指標としての効果を把握するために売上との相関, および売上よりも時間的な先行性があるかを確認することが重要である. また, それが満たされる場合需要予測モデルの一要素としての効力を発揮すると考えられる. そのため, 今回は Blog 情報の先行指標としての効用性を探ることにした. 本研究では一例として芥川受賞というイベントに着目し, 芥川賞受賞前後の作品の売上データおよび該当作品の Blog での言及数, およびテキスト情報の解析を行い,Blog 情報の売上先行指標としての効力を測る事にした. 芥川賞を選んだ理由は, 賞の候補, 受賞などが行われる時期が事前に把握できており, それらの影響力が Blog により補足されていると考えられ, その効果を書籍の売上との相互相関を測る事により, イベントの売上への影響が確認できると考えられるからである.

3 書籍の Blog テキスト解析

3.1 対象書籍

今回利用したデータは 2005 年度から 2009 年度までに芥川賞を受賞した 9 書籍である. 芥川賞は主に新人を対象とした賞である. 表 1 に該当書籍情報を記す.

表 1: 2005 年から 2009 年までの芥川賞受賞作品

受賞作	著者名	受賞年月日
土の中の子供	中村 文則	2005-7-13
沖で待つ	絲山 秋子	2006-1-16
八月の路上に捨てる	伊藤 たかみ	2006-7-12
ひとり日和	青山 七恵	2007-1-15
アサッテの人	諏訪 哲史	2007-7-16
乳と卵	川上 未映子	2008-1-15
時が滲む朝	楊 逸	2008-7-14
ポトスライムの舟	津村 記久子	2009-1-14
終の住処	磯崎 憲一郎	2009-7-14

3.2 Blog テキストデータ

今回利用した Blog テキストデータは, 芥川賞受賞書籍のタイトルをクエリとし,Yahoo!検索 Web API[1]において提供されているブログ検索 Web API を用いて取得したデータである. 本文のテキストデータは Yahoo Blog API で取得した該当 Blog URL から本文を抽出することで取得した. 表 2 にそれぞれの書籍において取得した Blog の数を記す.

表 2: 芥川賞受賞作品の Blog 取得数

クエリ	Blog 数
土の中の子供	112
沖で待つ	318
八月の路上に捨てる	154
ひとり日和	530
アサッテの人	457
乳と卵	970
時が滲む朝	412
ポトスライムの舟	804
終の住処	773

3.3 Blog テキストの形態素解析

Blog テキストの解析目的としては, 芥川賞・直木賞を受賞した作品に言及した Blog においてよく使われる語句を解析することにより, これらの書籍に共通で見られる特徴を掴むことにある. テキストの解析手法としては, 取得した Blog テキストに形態素解析を行い, 語句の頻度解析を行った. 形態素解析をするにあたっては MeCab[2]を利用した. 形態素解析をするにあたり, 書籍の内容および評価に関する形態素を抽出するために利用する品詞を名詞, 動詞, 形容詞に絞り解析を行った. また品詞細分類の中では非自立, サ変接続, 接尾, 代名詞, 数に該当する形態素は除外した. 表 3 に形態素解析統計データを記す. また, 表 4 に各 Blog 毎の形態素解析において上位に現れた形態素, および頻度を記す.

表 3: 取得した Blog の形態素解析統計データ

クエリ	形態素タイプ数	形態素の総数
土の中の子供	7272	18955
沖で待つ	15353	57578
八月の路上に捨てる	10600	28085
ひとり日和	19423	82355
アサッテの人	19636	84461
乳と卵	30641	169314
時が滲む朝	19516	84741
ポトスライムの舟	24706	161148
終の住処	28174	129850

[1]http://developer.yahoo.co.jp/webapi/search/

[2]http://mecab.sourceforge.net/

表 4: 取得した Blog の上位 30 形態素および頻度

土の中の子供		沖で待つ		八月の路上に捨てる	
する	997	する	2220	する	1353
ある	319	半身	2141	ある	388
読む	294	ある	668	読む	353
なる	263	なる	610	なる	350
思う	220	読む	600	思う	307
ない	210	ない	467	ない	288
作品	184	思う	406	作品	241
子供	161	沖	355	人	177
自分	151	待つ	350	本	174
芥川賞	142	本	306	月	174
中村	140	作品	291	自分	172
人	131	人	282	捨てる	169
本	126	絲山	277	芥川賞	168
小説	126	芥川賞	260	八月	155
土	118	秋子	248	路上	149
主人公	110	言う	248	伊藤	145
言う	105	月	238	たかみ	132
文則	104	小説	232	小説	124
いう	98	文庫	193	言う	105
書く	94	いう	181	できる	100
下半期	80	自分	172	見る	94
上半期	80	見る	166	いう	93
大江	76	いい	157	上半期	90
人間	76	いる	155	いる	87
月	70	できる	147	下半期	85
できる	67	書く	139	世界	78
生きる	66	女性	132	敦	76
感じる	64	太る	126	書く	76
いる	63	出る	126	感じ	70
いい	61	物語	125	感じる	70

ひとり日和		アサッテの人		乳と卵	
する	4098	する	4506	する	8530
なる	1098	ある	1121	ある	2193
ある	1040	読む	1094	読む	2162
読む	897	なる	1034	なる	2083
文庫	871	人	993	思う	1561
ない	751	小説	768	川上	1519
思う	664	思う	731	ない	1345
月	543	作品	727	映子	1226
作品	475	ない	719	月	1200
人	473	アサッテ	702	卵	1164
ひとり	436	芥川賞	586	芥川賞	1150
本	394	書く	507	乳	1101
ライブ	388	月	451	作品	980
日和	382	本	424	書く	909
芥川賞	380	諏訪	417	小説	870
小説	378	いう	376	本	842
青山	351	言葉	375	人	761
自分	331	文学	357	言う	732
いう	324	言う	342	いう	685
見る	298	哲史	338	自分	593
七恵	284	見る	323	できる	580
文学	282	できる	313	見る	559
前	280	叔父	312	文学	556
いる	277	物語	267	世界	548
ブログ	276	世界	267	いる	486
書く	270	直木賞	256	作家	473
言う	264	自分	252	女性	464
できる	260	いる	245	文章	408
日本	231	面白い	202	いい	404
主人公	221	いい	198	大阪	402

表 4 から分かることとしては「時が滲む朝」の上位 6 位, 19 位形態素にそれぞれ「中国」,「中国人」という形態素がある事から, 中国人が芥川賞を取得したことが注目すべきニュースとして Blog に書かれた様子が伺える. また, 女性著者が芥川賞を受賞した際には「女性」という形態素が上位に来る様子が伺える. 実際,「沖で待つ」,「乳と卵」,「ポストスライムの舟」の上位 30 形態素に「女性」という形態素が入っている事から女性著者が芥川賞を受賞した事が Blog 上で話題になった事が推測される. 評価を表す形態素としては「いい」,「面白い」という要素が 20 から 30 位の上位形態素として出現しているものが見られた. 評価に関わる形態素は最上位の 1 から 20 位以内にはないが 20 位以降の要素として出現する事が多く, また, 評価としては好意的な評価が多い様子が伺える.

4 書籍の売上と Blog での言及数の解析

4.1 書籍の売上データ

今回利用した書籍売上データは書籍取次会社から提供された書籍の売上データを用いた.

4.2 書籍売上データ Blog ヒット数との相互相関解析

書籍売上および, 書籍の Blog での言及数に関する相互相関解析を行う. 相互相関解析を求めるのは, Blog が

時が滲む朝		ポトスライムの舟		終の住処	
する	4188	する	9192	する	6563
なる	1074	月	2447	ある	1768
ある	1068	ある	1973	なる	1724
読む	838	なる	1847	読む	1247
芥川賞	784	人	1283	思う	1085
中国	676	思う	953	月	993
日本	665	ない	949	ない	872
思う	642	読む	948	作品	730
作品	628	できる	689	終	688
ない	507	芥川賞	606	住処	675
朝	492	平成	599	人	629
月	482	http	589	芥川賞	564
小説	480	社会	566	書く	528
書く	463	いう	558	小説	525
滲む	438	自分	550	言う	523
日本語	437	ニュース	545	本	504
文学	428	作品	538	自分	462
楊逸	414	jp	531	見る	457
中国人	395	ポトス	483	できる	452
人	338	東京	476	磯崎	449
言う	338	女性	475	いる	434
事件	335	いる	470	いう	431
本	327	トップ	447	妻	363
天安門	310	考える	443	時間	357
作家	295	問題	442	今	345
直木賞	294	ライム	440	日本	341
いう	278	言う	440	憲一郎	317
できる	265	本	434	著者	312
見る	252	精神	421	前	309
浩	231	多い	418	人生	304

書籍売上データに対して先行指標となるか確認を行い, 相関がある場合売上データとラグはどれほどあるか確認するためである. 相互相関を求めるにあたって利用した期間は芥川賞の受賞日の 30 日前を起点とした 90 日分のデータを利用した. 受賞日の 30 日を起点としたのは,Blog 情報が売上情報に先行するか確認するためである. また, 相互相関の Lag 範囲は 0 日から 30 日の範囲で計算を行った.

表 5 に芥川賞の受賞日の 30 日前を起点とした 90 日分の書籍の売上および Blog 取得数を記す. 表 5 から分かることとしては,「土の中の子供」,「沖で待つ」,「八月の路上に捨てる」,「ひとり日和」の Blog 数は全て 30 未満という非常に少ないデータである事が分かる. この期間に取得できた Blog 数が少ない理由としては,2005 年度から 2007 年前半までは Blog の絶対数がそもそも少ない, もしくは Blog のクローリングを本格的に行なっていなかったという事が理由として考えられる. また, 2007 年度上半期に芥川賞を取得した「アサッテの人」を境に Blog の絶対数が大幅に増え, それ以降のデータも安定した Blog の数を取得出来ている事からこの時期を境に比較的安定的に Blog を活用した人が増えた, また Blog のクローリングシステムが本格稼働し出したと推測できる.

表 5: 芥川賞受賞作の受賞期間近隣 90 日範囲の書籍売上数および Blog 取得数

クエリ	売上数	Blog 取得数
土の中の子供	2386	14
沖で待つ	2129	19
八月の路上に捨てる	1484	12
ひとり日和	3894	28
アサッテの人	2461	225
乳と卵	4705	463
時が滲む朝	4391	267
ポトスライムの舟	3531	254
終の住処	8217	287

表 6 に相互相関係数の解析結果を記す. また図 1 に「ポトスライムの舟」, 図 2 に「終の住処」の相互相関係数図を記す.

表 6: 書籍の売上と Blog ヒット数の最大相互相関係数

書籍	最大相互相関係数	Lag(日)
土の中の子供	0.234	20
沖で待つ	0.302	3
八月の路上に捨てる	0.117	5
ひとり日和	0.309	4
アサッテの人	0.446	4
乳と卵	0.537	6
時が滲む朝	0.374	1
ポトスライムの舟	0.397	2
終の住処	0.365	8

表 6 から分かる事としては, 「八月の路上に捨てる」の 0.117 という値を除くと, 他の全ての書籍において 0.3 以上の中程度以上の相互相関係数値を記録している事が分かる. また, 最大相互相関係数を記録した Lag の日数も 1 から 20 日までと幅はあるものの,「終の住処」の 8 日や,「土の中の子供」の 20 日のように一週間以上先行して最大相互相関値を記録したものの存在が確認できた. この事から芥川賞というイベントを介した際の Blog での書籍言及数が売上に先行して現われる事が推測される.

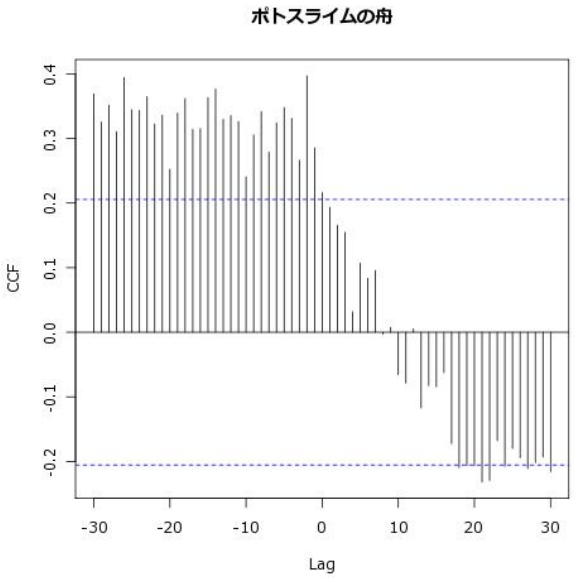

図 1: ポトスライムの舟の相互相関係数

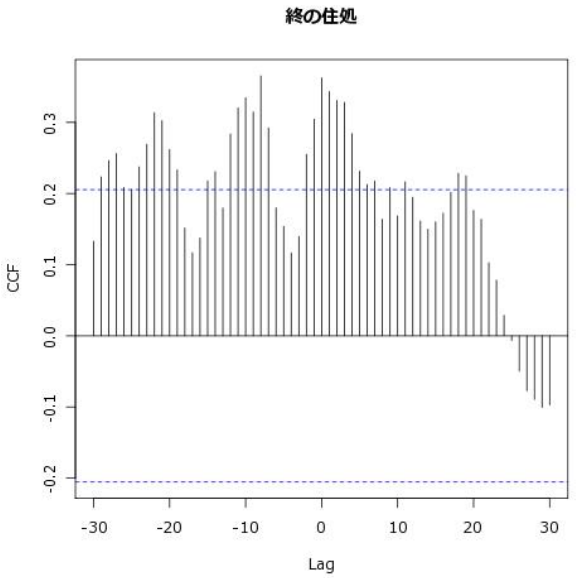

図 2: 終の住処の相互相関係数

図 1 からは, 比較的長期期間において高い相互相関係数の値を記録している事から, 長期期間の間 Blog が書籍の先行指標となることを示唆している. 図 2 からは, 相互相関係数の値が 12 日程度の周期で高い値がでる事から, 周期的なイベントなどの要因が売上と書籍の売上に影響を及ぼしたと考えられる.

4.3 書籍売上データと Blog テキストデータを利用した相互相関解析

Blog での単純な言及数に Blog テキストのポジネガ分析を重みとして付け加えた際に, その重みと書籍売上との相互相関係数がどのように変化するかの実験を行った.

この実験の目的は,Blog 上での書籍の評価が売上に影響を与えるという仮定の元に, 書籍に対する評価という重みを単純な Blog ヒット数に付け加えることで, その効果があるか調べる事にある.

Blog テキストのポジネガ分析を重みづけするにあたって, 単語感情極性対応辞書 [10] を利用することにした. この辞書は日本語の形態素のうち, ポジティブ, もしくはネガティブと判定された形態素を数値で重み付けをし, 重みの範囲を-1 以上 1 以下の値に調整してある. 重みの計算式は以下の通りである. この数式において,L は Blog に含まれる形態素集合, ω は形態素, $c(\omega)$ は形態素 ω のポジネガの重み, N は形態素の総数を表している. この数式では各々の Blog の重みの範囲は 0 以上 2 以下に収まる.

$$\alpha = 1 + \sum_{\omega \in L} \frac{c(\omega)}{N} \tag{1}$$

表 7 に書籍の売上と Blog の言及数に Blog テキストのポジネガの重み付けをした相互相関係数の解析結果を記す.

表 7: 書籍の売上と Blog ヒット数にテキストの重み付けを加えた指標の最大相互相関係数

書籍	最大相互相関係数	Lag(日)
土の中の子供	0.246	19
沖で待つ	0.325	2
八月の路上に捨てる	0.151	4
ひとり日和	0.368	2
アサッテの人	0.357	12
乳と卵	0.302	9
時が滲む朝	0.249	22
ポトスライムの舟	0.352	29
終の住処	0.304	7

表 6 と表 7 を比較することで, テキストの重み付けの結果の解析を行う. まず,「土の中の子供」,「沖で待つ」,「八月の路上に捨てる」,「ひとり日和」はそれぞれ最大相互相関係数が上昇している様子が分かる. その中でも「ひとり日和」は 0.059 と比較的大きく相互相関係数の値が上がっていた事が分かった. また, Lag も絶対値 2 日以内に収まっていたことから, テキストの重み付けの効果があったと推測される. この 4 つの書

籍は, そもそも Blog の絶対数が少ないという特徴をもつ. Blog の絶対数が少ないことにより, ノイズとなるような Blog 情報が少なかったためテキストの重み付けが効力を出したのだと推測できる.

次に,「アサッテの人」,「乳と卵」,「時が滲む朝」,「ポトスライムの舟」,「終の住処」はそれぞれ最大相互相関係数が減少して, 減少が大きいものとしては「乳と卵」が 0.235 も値を下げた. 最大相互相関係数を記録した Lag(日) も大きく変動した書籍が多く,「ボトスライムの舟」は Log の最大値が 2 日から 29 日へと変化している.

これら 5 つの書籍は Blog の取得数が 200 以上と比較的安定した量を取得しているのであるが, Blog の数が多くなるにつれノイズとなるような書籍の評価と関係がない Blog 情報の混入も多くなったと予測される. 書籍と評価と関係のない Blog が多く混入することにより, 重みそのものに影響を与え, 最大相互相関係数値に影響を与えたと考えられる.

5 結論と今後の展望

本研究では芥川賞というイベントに着目し 2005 年度から 2009 年度までの芥川賞受賞作品における書籍売上と, 書籍タイトルをクエリ情報として取得した Blog 情報の解析を行った.

その結果として以下のような事が分かった. 第一に Blog テキストの形態素解析結果としては, 芥川賞を受賞した作品の社会的な意味合いの強い形態素が Blog を書く際に併記される事が多い事が分かった. 第二に Blog が売上に先行するかの調査をするために行った売上と Blog ヒット数との相互相関解析では Blog の取得数が極端に少ない場合は, 先行指標としての効力が少なく, 先行指標として Blog を使う場合には該当書籍に言及している Blog がある程度の数が必要である事が推測された. Blog の重みとしてテキストの感情評価を導入した解析では, 相互相関係数値が Blog 数が少ない際には上昇, Blog 数が多い場合には減少したが今回は書籍の絶対数が少なく, 感情評価の重みのつけ方も適用した手法が一種類と少なかったので, Yang らの [7] 手法などを試すなど, 更なる調査が必要である.

今後の展望としては, 本研究において Blog の書籍需要予測をする上での先行指標性の有用性が見出せたので, 実際に書籍売上数と Blog 情報を組み合わせた需要予測モデルを作成し, その効果を検証する事が考えられる.

参考文献

[1] 2008 出版指標年報, 出版法人全国出版協会　出版科学研究所, pp.3-5, 2008.

[2] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins.: Structure and evolution of blogspace. *Commun. ACM, 47(12)* , pp.35-39, 2004.

[3] Rubicon Consulting Inc. Online communities and their impact on business: ignore at your peril, October 2008.

[4] 文健哲，菊田剛，寺野隆雄：国内書籍販売の先行指標としての Blog 情報の活用. *Direct Marketing Review*, Vol. 9, pp. 33-48, 2010.

[5] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, Andrew Tomkins.: The predictive power of online chatter. *In KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 78-87, 2005.

[6] Wolfgang Jank, Galit Shmueli, and Shanshan Wang.: Dynamic,real-time forecasting of online auctions via functional models. *In KDD '06* , pp.580-585, 2006.

[7] Yang Liu , Xiangji Huang , Aijun An , Xiaohui Yu.: ARSA: a sentiment-aware model for predicting sales performance using blogs, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* , pp. 607-614, 2007.

[8] Yang Liu , Xiangji Huang , Aijun An , Xiaohui Yu.: S-PLASA+: adaptive sentiment analysis with application to sales performance prediction, *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval* , pp. 873-874, 2010.

[9] Thomas Hofmann. Probabilistic latent semantic analysis. *In UAI'99* , 1999.

[10] Hiroya Takamura, Takashi Inui, Manabu Okumura.: Extracting Semantic Orientations of Words using Spin Model., *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)* , pages 133–140, 2005.