2A3-2

データマイニング手法を利用した書籍POSデータの解析

Data Mining Approach to Book Sales Analysis Using Point of Sales Data in Japan

菊田 剛[*1] 文 健哲[*1] 山田 隆志[*1] 吉川 厚[*1] 寺野 隆雄[*1]
Go Kikuta Geun Chol Moon Takashi Yamada Atsushi Yoshikawa Takao Terano

[*1]東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institue of Technology

This paper presents a Book sales tendency by analyzing the point of sales data collected from 100 high-ranking sales books 2009 in Japan. We analyzed the tendency of Best-Seller Books, and the correlation with the number of references in sales and Blog. As a result, We found interesting characteristics in Best-Seller Books.

1. はじめに

近年, 出版業界は市場規模が縮小しており非常に厳しい状況が続いている. 市場規模の減少の原因としては, 少子高齢化, インターネットや携帯電話の普及による人々のライフスタイルの変化など様々な原因が考えられる. また, 今後普及するであろう電子書籍の影響による相対的な紙媒体の書籍の需要の減少なども予測され, 今後も厳しい状況が続くと考えられる.

そのような厳しい状況にも関わらず書籍の返品率は依然として高い状況が続いている. 返品率の高さの原因としては色々な要因が考えられるが, その中でも出版業界特有のは再販価格制度と委託制度という制度に依る所が大きいと考えられる. 再販価格制度は書店側が書籍の値下げが出来ないというもので, その結果在庫を抱えていても値下げをして売りさばくという事は出来ない. また, 委託制度は書店は販売の委託をされているのであり, 書店が本を購入して販売しているわけではないという制度である. その結果, 書店側は品切れによる機会損失を減らすために安易に過剰在庫策を採用する傾向がある. この結果, 図 1 に示すように返品率は常に 40%程度になっている [1].

書籍売上動向の適正な予測がつくようになると, 適正な需要の把握が可能になり, その結果書籍の返品率も下がると考えられる. 本研究では 2009 年度の上位ベストセラー 100 書籍の売上の解析, 及び売上と Blog での言及数との相関調査を行った. ベストセラー特有の傾向を掴むことで, 今後どのような本がベストセラーになるか事前に把握をする事が可能になると考えられる.

2. 関連研究

我々は過去の研究 [2] 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.

日本における書籍販売市場の法則性の研究としては伊庭ら [3] がある. この研究では書籍の販売冊数と順位の関係がべき乗分布に従っている事を示している. また, ジャンル毎の分布特徴があることも言及している.

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 [4] がある. この研究では Amazon[*1]における 2340 冊の書籍の 2004 年 7 月から 2004 年 10 月までのセールスランキングデータと Blog 情報とを利用し以下の三項目にわたる研究を行っている. 第一に適切に生成したクエリから取得した Blog 情報と Amazon のセールスランキングとの相関についての解析, 第二にクエリの自動生成方法, 第三に需要予測システムの開発である. Blog 情報と数理モデルを用いた需要予測に関する研究としては吉田ら [5] がある. この研究の対象は映画であり,Bass モデル [6] に Blog の口コミ情報を加えた数式モデルを立てている. また,Blog 情報のポジネガ分析を行いそれらの情報と興行収入との関わりについて解析を行っている.

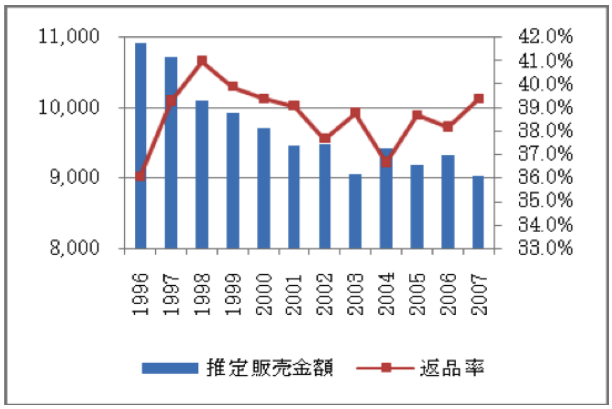

図 1: 書籍販売と返品率の推移 [1]

このような背景を踏まえ, 本研究では 2009 年度の単行本・新書売上ベスト 100 の全体解析, 及び Blog での言及数を踏まえた上位書籍の個別解析を行う事にした.

3. ベストセラー本の解析

3.1 利用データ

今回利用した書籍売上数データは書籍取次会社から提供されたデータを用いた. 書籍数は 2009 年度の売上上位 100 冊, ランキングの基準は紀伊国屋書店が発表しているキノベス'09[*2]. の単行本・新書売上ベスト 100 を指標とした. 期間は 2009/1/1 から 2009/12/31 までの 365 日, 書店数は 2477 書店である. 注意事項としては, 指標として利用したデータと今回抽出したデータのランキングに関しては必ずしも一致していない点があ

連絡先: 東京工業大学大学院総合理工学研究科, 〒 226-8502 横浜市緑区長津田町 4259 J2 棟 1704

*1 http://www.amazon.com

*2 http://www.kinokuniya.co.jp/01f/kinobes/2009/best100.htm

る事があげられる.

3.2 全体傾向の解析

表 1 に 09 年度のベストセラー本の売上冊数の概要, 図 2 に上位売上数から始まるベストセラー本の片対数グラフを記す. 表 1, 図 2 から分かることとして, 平均とメディアンとの差は 0.26 σ と非常に大きく, ランキングの分布が下方に集中している事が把握できる. また, 平均値と最大値の差は 5.28 σ である. これはランキング上位と下位の差が非常に大きい事を示している. 実際ベストセラー 100 書籍の総売上冊数のうち 10%が上位 2 書籍,20%が上位 6 書籍によって占められる結果となった. ベストセラーランキングであるにも関わらず, 最小値が極端に少ないのは, 紀伊国屋とは違い, 取次の提携書店ではこの書籍の売れ行きが芳しくなかった事を意味する.

表 1: ベストセラー本の書籍別売上データ

最小	メディアン	平均	最大	標準偏差
49	10890	14140	80230	12518

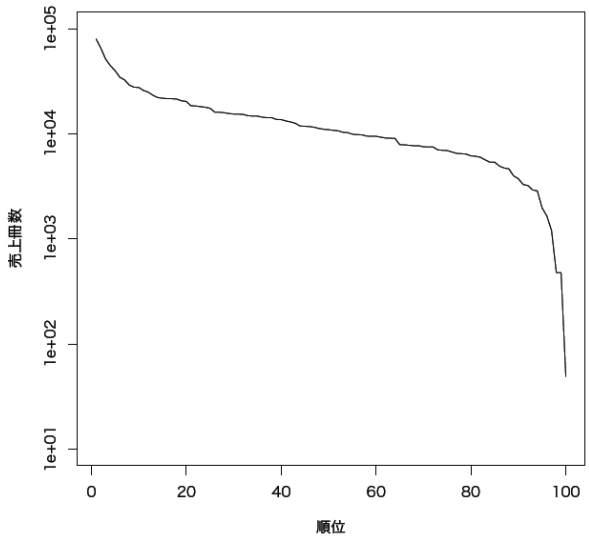

図 2: ベストセラー本の書籍別売上グラフ

表 2: 日単位のベストセラー総合売上データ

最小	メディアン	平均	最大	標準偏差
651	3709	3873	10810	1473

次に, 図 3 に日単位のベストセラー総合売上動向, 表 2 に統計データを記す. 図 3, 及び表 2 から分かることとしては日毎における書籍売上の標準偏差が 1473 と大きく, 日単位での書籍売上数の分散が大きいことが把握出来る.

また, 一日の売上が 10,000 を越えた日は 5/29,30 日の 2 日であり, この日は 2009 年度の年間ベストセラー一位を獲得した 1Q84 の 1, 2 巻の発売日である事を考慮する事で, ベストセラー本の発売日が一日における書籍の総売上に大きな影響を及ぼしている事が推測できる.

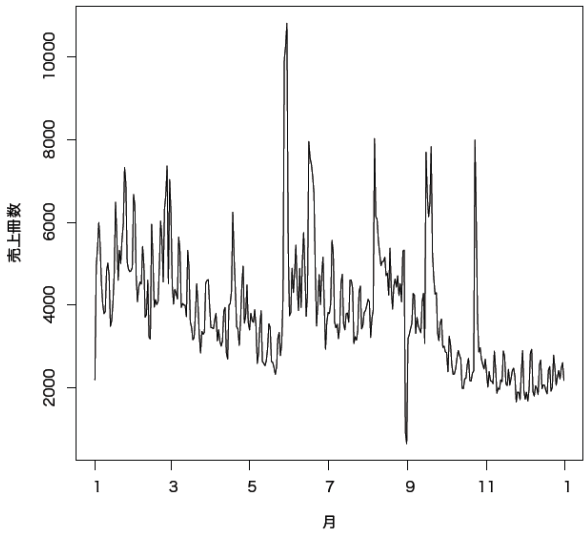

図 3: 日単位のベストセラー総合売上グラフ

3.3 個別書籍の解析

個別書籍の解析をするに当たり, 今回はベストセラー 100 書籍において売上数上位 20%を占める上位 6 書籍を利用した. 解析内容は書籍の売上数と Blog における言及との相関解析である.

Blog 情報を取得するに当たり, 今回は Yahoo! Blog 検索[*3]を利用した. 検索クエリには各書籍のタイトルを利用し, 取得データは各日付毎のクエリのヒット数とした. Blog のヒット数取得範囲は 09 年度における 1 年スパン, 書籍発売月の前月を起点とした 4ヶ月スパン, 及び発売日の前月から発売月までの 2ヶ月スパンの三種類である.

表 3 に対象書籍の売上数を記す. 売上のランキングの基準はキノベスを基準としているので, 提供されたデータの売りげと前後する部分がある.

表 4 に対象書籍の Blog ヒット数を記す. 表 4 から分かることとしては,「1Q84」のように新聞やニュースで話題になった書籍は発売月の 1ヶ月前を起点とした 4ヶ月スパンにおける Blog での言及数が高い事, また, ベストセラーにも関わらず「読めそうで読めない間違いやすい漢字」や「バンド 1 本でやせる！巻くだけダイエット」のように発売日の前後では Blog 上で全く話題とならなかった事が確認された. また, 発売日の全月を起点とした 2ヶ月スパンと 4ヶ月スパンとの Blog ヒット数を比較すると, 全体傾向として 4ヶ月スパンでの Blog ヒット数は,2ヶ月スパンの 2 倍以上のヒット数が出る傾向にある事が分かった. 例えば「1Q84 Book1」は 2ヶ月スパンと 4ヶ月スパンで実に 6.7 倍もの差がある. この事から, 書籍に関する Blog での言及は発売日前には傾向的に少なく, 書籍の発売後その本を読んだ読書が Blog に読書感想を書き記したりする事により言及数が上がっていくと推測される.

表 5 に年間書籍売上数と年間 Blog ヒット数とのピアソンの積率相関係数を示す. 表 5 から分かることとしては, 「1Q84」のように同時期に発売された相互に関連する書籍の相関値は比較的似た値を取ることが確認された. また, ランキング上位の本だからといって売上と Blog との相関が見られないものが

*3 http://blog-search.yahoo.co.jp

表 3: キノベス上位 6 書籍の総売上数

書籍名	売上数
１Ｑ８４　ＢＯＯＫ１	80229
読めそうで読めない間違いやすい漢字	51738
１Ｑ８４　ＢＯＯＫ２	65505
「脳にいいこと」だけをやりなさい！	34671
オバマ演説集	18525
バンド 1 本でやせる！巻くだけダイエット	44716

表 4: キノベス上位 6 書籍の Blog ヒット数

ランキング	年間	4 ヶ月	2 ヶ月
1	1222	795	118
2	296	1	1
3	992	657	92
4	446	181	29
5	497	449	25
6	133	7	2

存在している事が分かる. 相関が見られなかった「読めそうで読めない間違いやすい漢字」や「オバマ演説集」は両者共に学習本であり, 「「脳にいいこと」だけをやりなさい！」は自己啓発本である. また, 中 高程度の相関が見られた「1Q84」や「バンド 1 本でやせる！巻くだけダイエット」はそれぞれ日本文学, ダイエット本である. この事から, 書籍のジャンルが書籍の売上と Blog での言及数との相関に何かしらの影響を及ぼしていると推測される.

表 5: キノベス上位 6 書籍の売上と Blog ヒット数との相関係数

書籍名	相関
１Ｑ８４　ＢＯＯＫ１	0.70
読めそうで読めない間違いやすい漢字	-0.09
１Ｑ８４　ＢＯＯＫ２	0.73
「脳にいいこと」だけをやりなさい！	-0.21
オバマ演説集	-0.14
バンド 1 本でやせる！巻くだけダイエット	0.52

次に, ピアソンの積率相関係数で中 高程度の相関が見られた「1Q84 BOOK1」,「1Q84 BOOK2」, の相互相関解析の結果を図 4, 図 5 に記す. この 2 書籍は同時期に出版されたシリーズ本であり, 年間売上の両者の相関が 0.99 と非常に高いのが特徴である. 図 4, 図 5 から分かることとしては, 相互相関のピークが Lag0 日付近であることから, 売上と Blog での言及数の推移が同時期である事, また, 売上及び Blog での言及数が違うにも関わらず両者の CCF のグラフは非常に似通っている事が分かる. この事から,「1Q84」のように同時期に出版されたシリーズ本においても似たようなグラフの傾向が出てくると推測される.

4. 結論と今後の展望

本研究では 2009 年度の単行本・新書売上数ベスト 100 書籍の統計解析, 及びその中でも売上の 20%を占める上位 6 書籍のにおける売上数と Blog ヒット数との相関に関する調査を行った.

その結果として以下のような事が分かった. まず, ベストセラーと言えども必ずしも Blog 上で言及数が多いわけでは無い

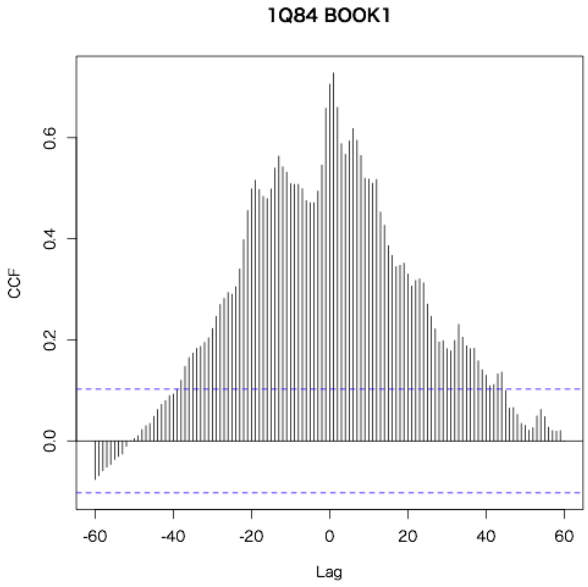

図 4: “1Q84 BOOK1” の相互相関係数グラフ

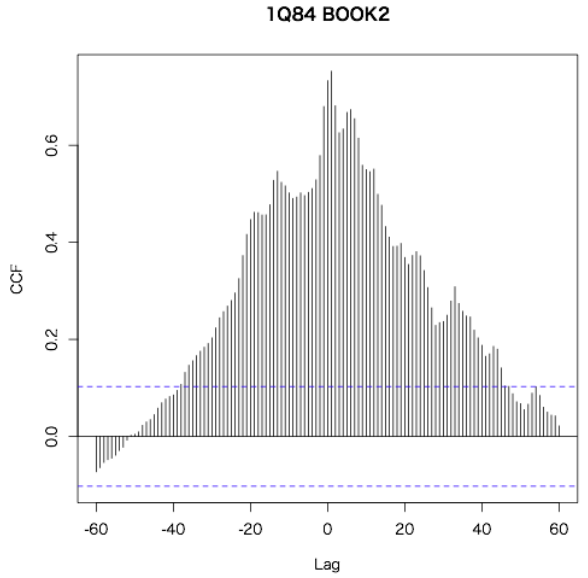

図 5: “1Q84 BOOK2” の相互相関係数グラフ

事が観測された. これは Blog を書く世代や Blog に言及されやすい書籍による影響が強いと考えられるので, 今後はそのような影響度を加味した調査が必要になると考えられる. また, 発売日以前での書籍に対する Blog での言及数が少ない事が把握された. この事から, 書籍の発売日以前に書籍情報が Blog において話題になる事は少ないと推測される. 今後は対象書籍を増やすことにより, より一般的な傾向を把握する事が重要になると考えられる. また, 書籍のジャンルごとに, 売上数と Blog との相関に差異が見られる事を確認した. 今回の解析は全 6 書籍と調べる数が少なかったので, 今後はジャンル別に書籍売上数と Blog での言及数との相関調査を行い, 売上と Blog との相関が強いジャンルの把握などを行う事が考えられる.

相互相関解析としては, 同時期に発売された 1Q84 の 1, 2 巻における, 売上と Blog との相互相関グラフ形状の把握を行った. その結果,1Q84 の 1 巻と 2 巻とでは売上において 15,000 冊ほどの差異,Blog ヒット数で 230 の差があったにも関わらず相互相関グラフの形状は似た形状をする事が判明した. この結果から,1Q84 のように同時期にリリースされた姉妹書においても同様な傾向があると推測できるので, この傾向が他の姉妹書にも当てはまるか調査を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] 文健哲，菊田剛，寺野隆雄：国内書籍販売の専攻指標としての Blog 情報の活用．*Direct Marketing Review*, Vol. 9, pp. 33-48, 2010.

[3] 井庭崇, 深見嘉明, 斉藤優: 書籍販売市場における隠れた法則性, 情報処理学会論文誌. 数理モデル化と応用, vol.48 SIG 6, 128-136, 2007.

[4] 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.

[5] 吉田就彦, 新垣久史, 石井晃, 林隆文, 梅村早苗: ヒット現象の数理モデル ～映画ヒットにおける Blog 分析～, 日本マーケティング・サイエンス学会 第 *83* 回研究大会, 2008.

[6] Frank M. Bass.: A new-product growth model for consumer durables. *Management Science*, Vol. 15, No.5, pp. 215-227, 1969.