Google Trendsを用いた検索の数理モデル

芦田昇 (鳥取大学 工学部)　石井晃 (鳥取大学 工学部)　川畑泰子 (群馬大学　社会情報学部)

1　序論

1.1　研究背景

Blog、Twitter、Facebook などのソーシャルメディアの普及が進み、インターネット上には様々な話題について書き込まれている。

著者の所属する研究グループではそのソーシャルメディアへの書き込みを用いた分析を行っている。分析を行うにあたって、鳥取大学の石井晃・デジタルハリウッド大学大学院の吉田就彦らによって提案されたヒット現象の数理モデルという数式を用いている [1]。過去の研究として、映画の興行収入の予測や AKB 選抜総選挙の順位予想などエンターテインメントの分野への応用に成功している [2]。従来は人々の興味関心の指標として Blog、Twitter を用いる数理モデルに対し、別の指標でも応用できないかと考え、本研究では検索数を人々の興味・関心の指標として計算する数理モデルを提案する。先行研究でも検索数を人々の興味関心の指標とし、SIR モデルを用いて情報の拡散過程を説明するモデルを構築していたが [3]、本研究で提案するモデルは検索行動の動力学を記述するモデルである。それにより、今までからは読み取れなかった潜在層の人々の動きや、及ぼされる外的影響から新たな知見が得られると考えられる。

1.2　研究目的

本研究はヒット現象の数理モデルにならう形で、社会の人々による検索行動の動力学を記述する新しい数理モデルを構築することを目的とする。その数理モデルでは人々の検索行動に、Blog や Twitter も影響を与えると考え、その影響とマスメディアの影響の比較も行う。本研究で特に注目していることは以下である。

1. 計算が実測の GoogleTrends を説明できるか、R factor から精度を確認、条件を変えて比較する。
2. 従来のモデルとパラメータを比較、考察する。
3. 新たに外力項に追加した Blog、Twitter のパラメータがどのように影響しているのかをみる。

なお、GoolgeTrends で用いている検索数のデータは 1 から 100 に規格化されたものであり、ここでは以下、関心度と呼ぶことにする。

2　ヒット現象の数理モデル

2.1　従来型ヒット現象の数理モデル

ヒット現象の数理モデルでは、社会における人々の 1 人 1 人が抱く興味・関心を定量化して方程式にしている。ある人「i さん」が抱く興味・関心を $I_i(t)$ と定義し、興味・関心（$I_i(t)$）を掻き立てる要員として、

1. メディアによる影響
2. 会話による影響
3. 噂による影響

の 3 つがあると考える。会話による影響のことを「直接コミュニケーション」とよび、街中の噂やソーシャルメディア上でのやり取りなどで影響を受けることを「間接コミュニケーション」と呼ぶことにする。それらについて興味・関心の時間的変化を追う微分方程式を立て、数理モデル化すると、以下の形で表せると仮定できる。

$$\frac{\mathrm{d}I_i(t)}{\mathrm{d}t} = \sum_{\xi} c_\xi A_\xi(t) + \sum_{j\neq i}^{N} D_{ij} I_j(t) + \sum_{j}\sum_{k} P_{ijk} I_j(t) I_k(t) \tag{1}$$

ここで D_{ij} は直接コミュニケーションの強さを表す係数、 P_{ijk} は間接コミュニケーションの強さを表す係数である. 広告宣伝の影響は外力と考える。ある話題に関する日毎のテレビでの露出秒数や、ネットニュースの件数を A_ξ とし、その係数を c_ξ とする。ξ はメディアの種類を表す添え字である。

式 (1) は個々の視点に基づいた式であるが、このままの形では分析を行うのは難しい。そこで、簡単化のために平均場近

似を行う。社会全体の構成員の数を N 人とし、社会全体で平均化された人々の意欲・関心を I(t) として、以下で定義する。

$$I(t) = \frac{1}{N}\sum_i I_i(t) \tag{2}$$

この平均場近似を用いると、 I(t) に従う方程式は次のようになる。

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \sum_\xi c_\xi A_\xi(t) + DI(t) + PI(t)^2 \tag{3}$$

導出の詳細は [1]、[4] を参照。
実際に計算する際には左辺の微分の箇所を、

$$\frac{\Delta I(t)}{\Delta t} = \sum_\xi c_\xi A_\xi(t) + DI(t) + PI(t)^2 \tag{4}$$

と表す。本研究では 1 日 1 日の動向をこのヒット現象の数理モデルで探るので、$\Delta t = 1$[日] として、

$$\Delta I(t) = \sum_\xi c_\xi A_\xi(t) + DI(t) + PI(t)^2 \tag{5}$$

となる。つまり実際にはこの $\Delta I(t)$ を計算している。

2.2 外力項に減衰のパラメータを追加

従来型のヒット現象の数理モデルの広告宣伝項は過去の影響を加味せず、その日その日のテレビや、ネットニュースの Blog などに与える影響しか考えていなかった。その影響の減衰力が指数関数的に減少すると推測し、その項を追加すると、計算する日を t_0 とすると、
式 3 の A(t) を $\sum_{i=0}^{N}\sum_\xi c_{\xi_i} A_{\xi_i}(t) exp(t_i - t_0)(\alpha_\xi)$ (α_ξ は減衰のパラメータ）と書き換えると、

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \sum_{i=0}^{N}\sum_\xi c_{\xi_i} A_{\xi_i}(t) exp(t_i - t_0) + DI(t) + PI(t)^2 \tag{6}$$

となる。N 日前の $t = t_N$ と表せる。

2.3 GoogleTrends を用いたヒット現象の数理モデル

本研究で用いるヒット現象の数理モデルを示す。

$$\begin{aligned}\frac{\mathrm{d}I(t)}{\mathrm{d}t} =& DI(t) + PI(t)^2 \\ &+ \sum_{i=0}^{N} c_{tvi} A_{tvi}(t) exp(t_i - t_0)\alpha_{tv} \\ &+ \sum_{i=0}^{N} c_{newsi} A_{newsi}(t) exp(t_i - t_0)\alpha_{news} \\ &+ \sum_{i=0}^{N} c_{blog_i} A_{blog_i}(t) exp(t_i - t_0)\alpha_{blog} \\ &+ \sum_{i=0}^{N} c_{twitteri} A_{twitteri}(t) exp(t_i - t_0)\alpha_{twitter}\end{aligned} \tag{7}$$

このモデルでは人々の興味関心の指標である I(t) を GoogleTrends の関心度とし、従来型のモデルの広告宣伝の項には Tv、ネットニュースの影響しか加味していなかったが、第 5 項の Blog の影響、第 6 項の Twitter の影響を追加した。
本研究ではこのヒット現象の数理モデルを元に分析を進める。

2.4 精度の計算方法

また、フィッテイングの際に精度を表す指標として R_factor を用いる [5]。本研究では以下の式を用いる。

$$R = \frac{\sum_i (f(i) - g(i))^2}{\sum_i ((f(i))^2 + (g(i))^2)} \tag{8}$$

ここで、本研究では $f(i)$ はターゲット (映画など）に対する関心度の件数、$g(i)$ はヒット現象の数理モデルによるシミュレーション結果の値を用いる。

R は、$0 \leqq R \leqq 1$ の範囲で、値が小さいほどフィッティングの精度が良いと言える。

3 分析方法

本研究は以下の流れで分析を行っている。

1. データの取得
   Blog、Twitter、5ch(2ch)、ネットニュース書き込み件数 (クチコミ@係長から取得 ホットリンク社より提供)
   Tv 露出秒数 (クチコミ@係長から取得 エムデータ社より提供)
   関心度 (GoogleTrends から取得)
2. 日毎の関心度件数の推移を数理モデルで再現（フィッティング）
3. 得られたパラメータの関係性を探る

分析区間を区切り、区間ごとにフィッティングを行いそれぞれでパラメータを算出する。本研究では 2016 年に大ヒットした映画の「君の名は。」、2017 年に一時ブームとなった「ハンドスピナー」に焦点を当てて分析する。
「君の名は。」の分析区間の区切り方を以下、表 1 に示す。分析区間は公開前 1 か月から公開終了日までとしている。
(公開日：2016/8/26)

表 1 「君の名は。」フィッティングの区間の区切り方

period1	2016/7/25〜8/25
period2	8/26〜9/25
period3	9/26〜11/9
period4	11/10〜2017/1/10
period5	1/11〜4/14

続いて、「ハンドスピナー」の分析区間の区切り方を以下、表 2 に示す。分析区間は Youtuber のセイキンが取り上げて日本のメディアに初めて取り上げられた 2016/3/28 から関心度が終息する 11/6 までとした。

表 2 「ハンドスピナー」フィッティングの区間の区切り方

period1	2017/3/28〜5/8
period2	5/9〜5/22
period3	5/23〜6/17
period4	6/18〜8/17
period5	8/18〜11/6

4 計算結果

4.1 「君の名は。」計算結果

図 1 に分析対象「君の名は。」のフィッテンググラフを示す。分析期間は period1 から period5 までの全体のものである。

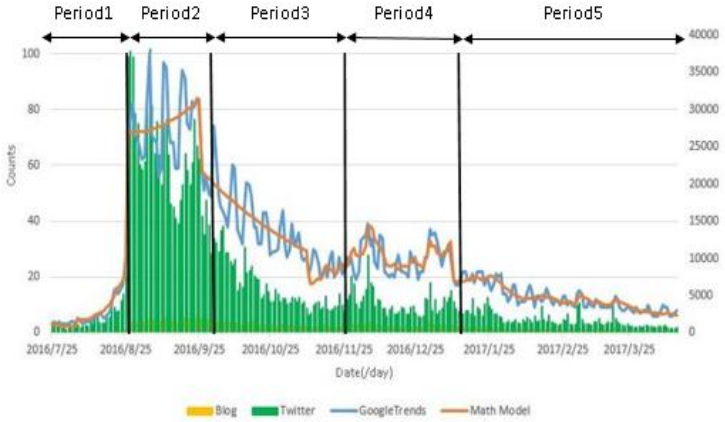

図 1　「君の名は。」フィッテング結果。：青線が関心度、赤線が数理モデルによる再現、緑の impulse が Twitter の件数、橙の impulse が Blog の件数である。横軸は 1 日おきの日付、縦軸は数値であり、左側を関心度のスケール、右側を Blog、Twitter の書き込み件数のスケールで表している。

次にパラメータを比較していく。以下に period1 から 5 までのそれぞれの直接コミュニケーション D、間接コミュニケーション P の比較グラフを図 2、図 3 に示す。

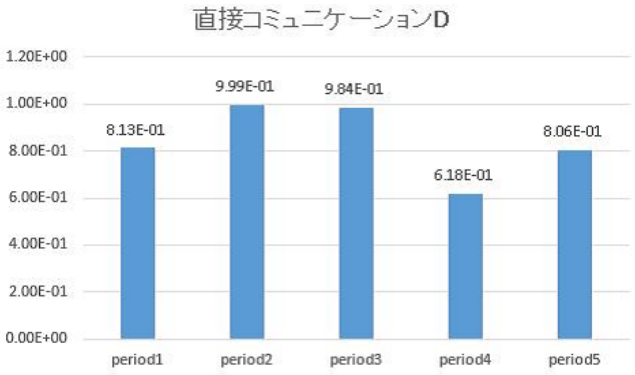

図 2　「君の名は。」における新型のモデルの直接コミュニケーション D 比較。横軸はそれぞれの分析区間、縦軸はパラメータの数値。分析区間に関しては表 1 参照。

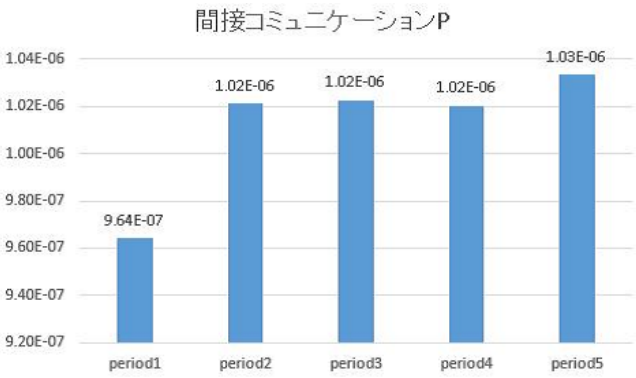

図 3　「君の名は。」における新型のモデルの間接コミュニケーション P 比較。横軸、縦軸、分析区間は図 2 と同じ。

同様に、Tv の影響 Cadv_t とネットニュースの影響 Cadv_n の比較グラフを図 4、図 5 に示す。

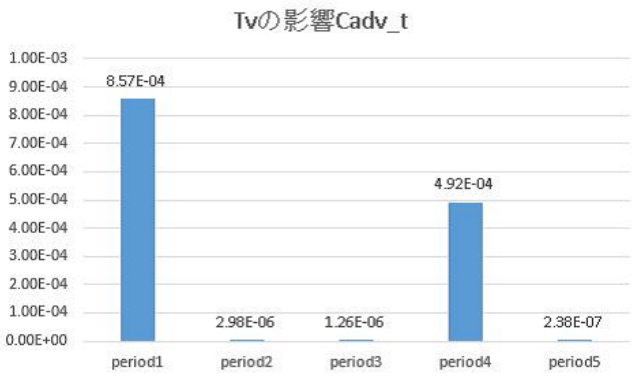

図 4　「君の名は。」における新型のモデルのテレビの影響 Cadv_t 比較。横軸、縦軸、分析区間は図 2 と同じ。

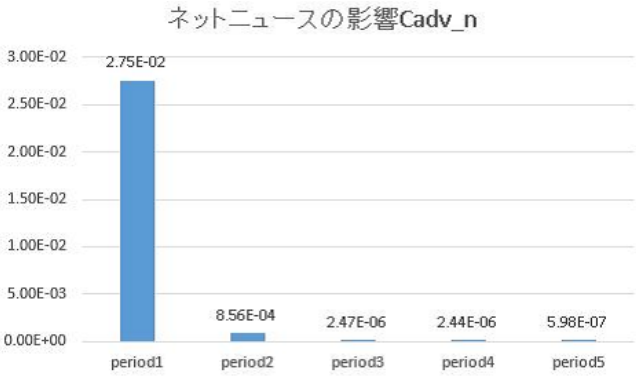

図 5　「君の名は。」における新型のモデルのネットニュースの影響 Cadv_n 比較。横軸、縦軸、分析区間は図 2 と同じ。

同様に、Blog の影響 Cadv_Blog と Twitter の影響 Cadv_twitter の比較グラフを図 6、図 7 に示す。

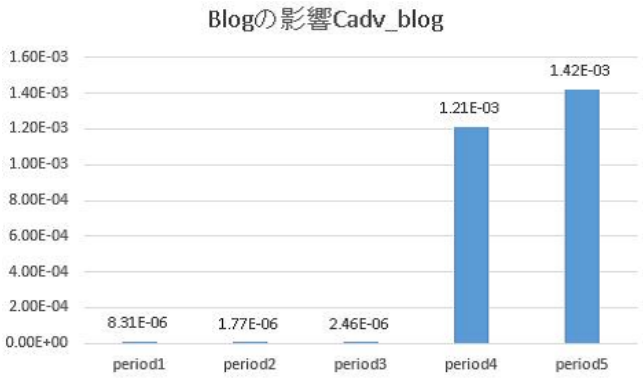

図 6　「君の名は。」における新型のモデルの Blog の影響 Cadv_blog 比較。横軸、縦軸、分析区間は図 2 と同じ。

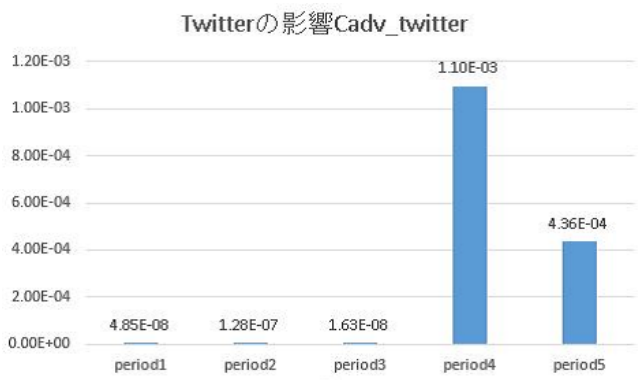

図 7　「君の名は。」における新型のモデルの Twitter の影響 Cadv_twitter 比較。横軸、縦軸、分析区間は図 2 と同じ。

次に、Twitter の書き込み件数を人々の興味・関心の指標とした従来型数理モデルでフィッテイングしたパラメータ D、P の比較グラフを図 8、図 9 に示す。

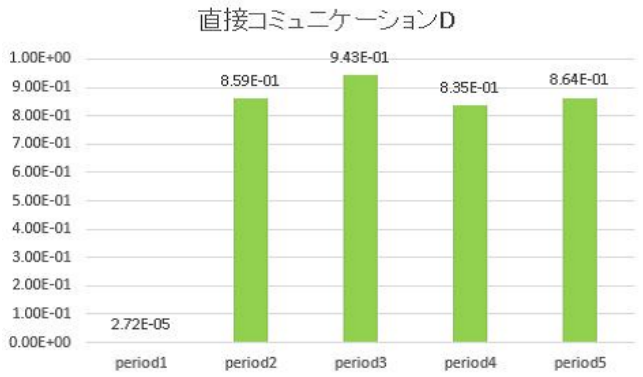

図 8　「君の名は。」における従来型モデルの直接コミュニケーション D の比較。横軸、縦軸、分析区間は図 2 と同じ。

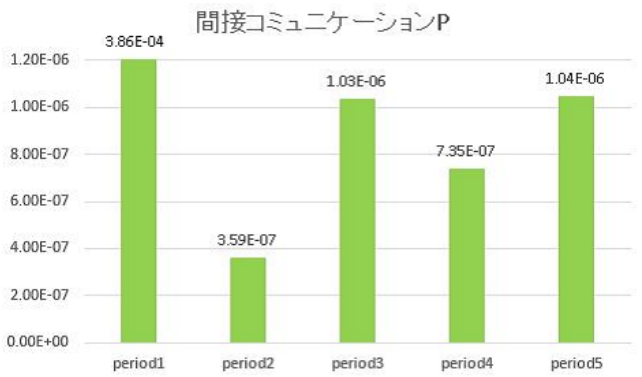

図 9　「君の名は。」における従来型モデルの間接コミュニケーション P の比較。横軸、縦軸、分析区間は図 2 と同じ。

4.2　「ハンドスピナー」計算結果

図 10 に分析対象「ハンドスピナー」のフィッテンググラフを示す。分析期間は period1 から period5 までの全体のものである。

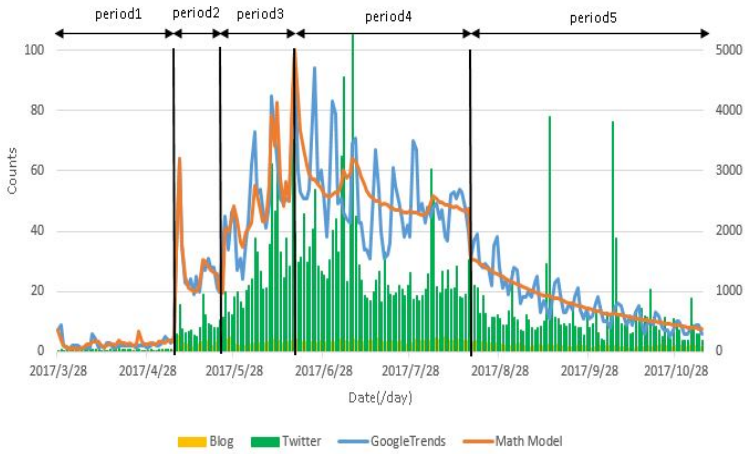

図 10　「ハンドスピナー」フィッテング結果。：青線が関心度、赤線が数理モデルによる再現、緑の impulse が Twitter の件数、橙の impulse が Blog の件数である。横軸は 1 日おきの日付、縦軸は数値であり、左側を関心度のスケール、右側を Blog、Twitter の書き込み件数のスケールで表している。

「君の名は。」と同様にパラメータを比較していく。以下に period1 から 5 までのそれぞれの直接コミュニケーション D、間接コミュニケーション P の比較グラフを図 11、図 12 に示す。

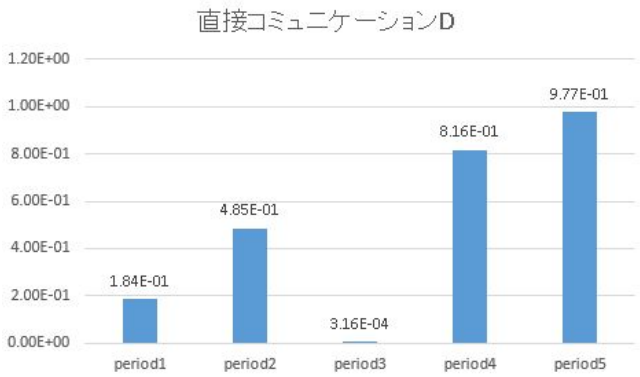

図 11　「ハンドスピナー」における新型のモデルの直接コミュニケーション D 比較。横軸はそれぞれの分析区間、縦軸はパラメータの数値。分析区間に関しては表 2 参照。

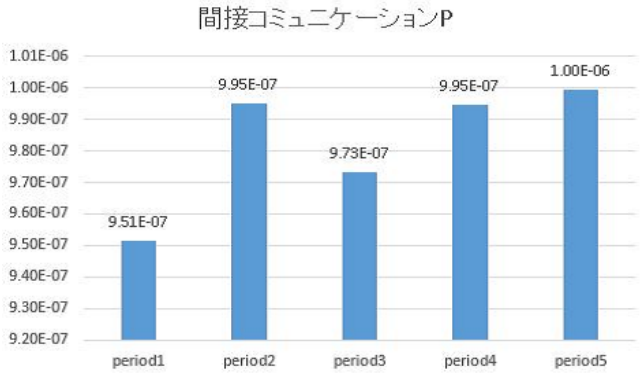

図 12　「ハンドスピナー」における新型のモデルの間接コミュニケーション P 比較。横軸、縦軸、分析区間は図 11 と同じ。

続いて、Tv の影響 Cadv_t とネットニュースの影響 Cadv_n の比較グラフを図 13、図 14 に示す。

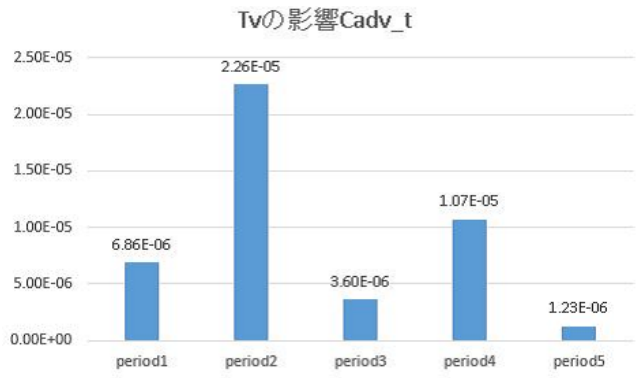

図 13　「ハンドスピナー」における新型のモデルの Tv の影響 Cadv_t 比較。横軸、縦軸、分析区間は図 11 と同じ。

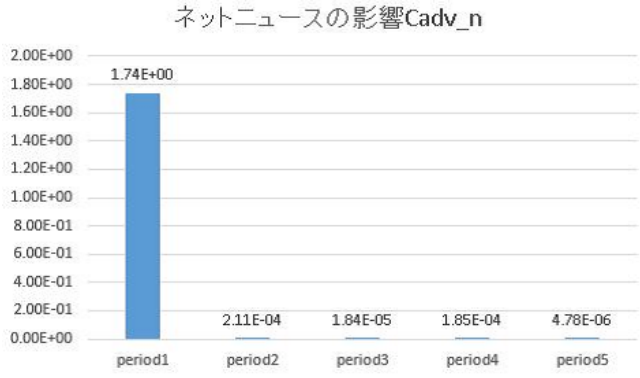

図 14　「ハンドスピナー」における新型のモデルのネットニュースの影響 Cadv_n 比較。横軸、縦軸、分析区間は図 11 と同じ。

次に、Blog の影響 Cadv_blog と Twitter の影響 Cadv_twitter の比較グラフを図 15、図 16 に示す。

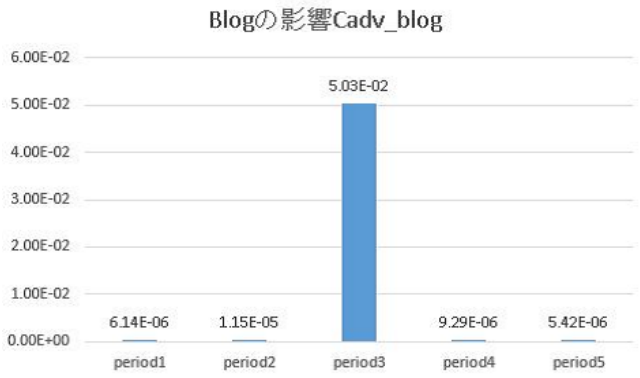

図 15　「ハンドスピナー」における新型のモデルの Blog の影響 Cadv_blog 比較。横軸、縦軸、分析区間は図 11 と同じ。

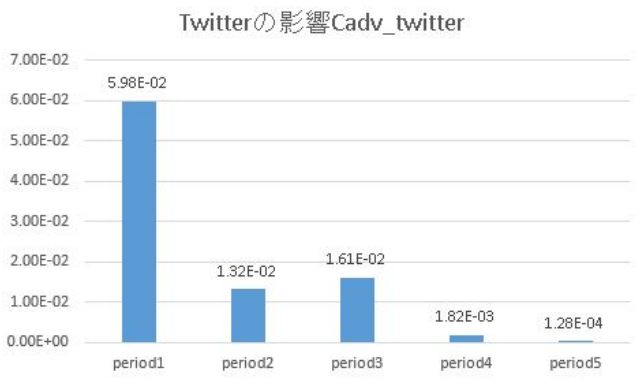

図 16　「ハンドスピナー」における新型のモデルの Twitter の影響 Cadv_n 比較。横軸、縦軸、分析区間は図 11 と同じ。

次に、Twitter の書き込み件数を人々の興味・関心の指標とした従来型数理モデルでフィッテイングしたパラメータ D、P の比較グラフを図 17、図 18 に示す。

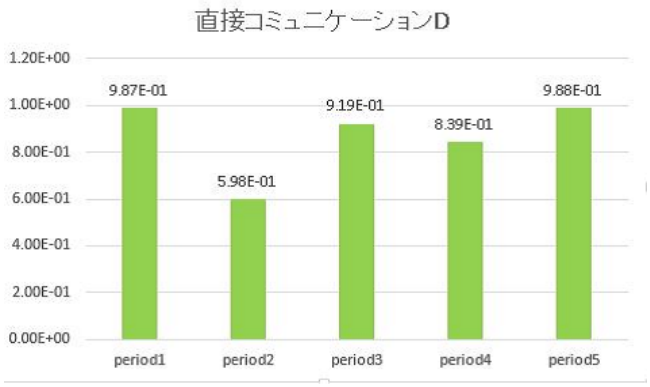

図 17　「ハンドスピナー」における従来型モデルの直接コミュニケーション D の比較。横軸、縦軸、分析区間は図 11 と同じ。

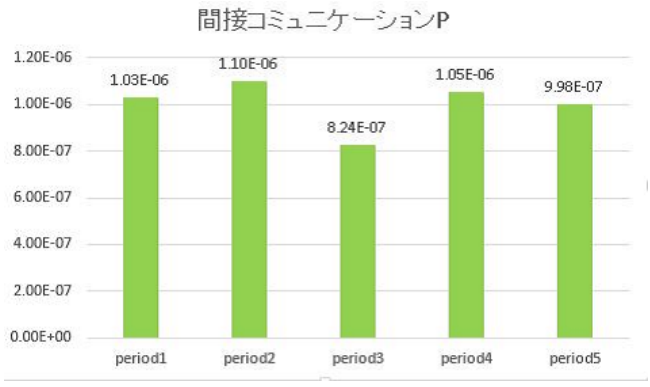

図 18　「ハンドスピナー」における従来型モデルの間接コミュニケーション P の比較。横軸、縦軸、分析区間は図 11 と同じ。

4.3　R_factor 計算結果

R_factor の計算結果を以下の表 3 に示す。人々の興味・関心の指標を Blog、Twitter にした場合と関心度にした場合、関心度は Blog、Twitter の影響を ON、OFF にした 4 パターンで比較した。分析区間は 1 か月、3 か月、6 か月を無作為にとった。結果として出た R_factor の平均値、中央値、最高値、最低値を表にしている。

表 3　R_factor の精度比較 (サンプル数 50)

興味・関心の指標	平均値	中央値	最高値	最低値
関心度 Blog,Twitter ON	0.031	0.026	0.119	0.001
関心度 Blog OFF	0.031	0.027	0.122	0.002
関心度 Twitter OFF	0.036	0.029	0.117	0.003
関心度 Blog,Twitter OFF	0.054	0.042	0.283	0.003
Blog	0.026	0.015	0.139	0.002
Twitter	0.097	0.075	0.32	0.007

5　考察

従来型の数理モデルでの D,P を顕在層の人々のパラメータ、本研究でのモデルの D,P を潜在層の人々のパラメータと仮定して考察を進める。それは Blog や Twitter が自分の考えを外に向けて発信していく人であるのに対し、GoogleTrends による検索行動には発信をしない潜在層の人々の関心も多く含まれているからである。

5.1　君の名は。

本研究で提案した数理モデルでの D が period4 以降減少しており、従来のヒット現象の数理モデルの D は公開後、一貫して高い値をとっている。このことから、顕在層の人々の間での話題にはなっている一方で、period4 以降の潜在層の人々が顕在層に移ったのではないかと考察できる。これは period4 以降での Twitter の影響が増大していることからも示唆できる。また、新型数理モデルの P が公開後高い値をとっていることからちょっとした会話にでていると言えるだろう。

次に、公開前の Tv、ネットニュースの影響が強まっており、検索行動に結びつく人々の興味関心は Tv、ネットニュースから影響を受けている。潜在層は Tv やネットニュースから情報を得ることから関心を得ていると言えるだろう。Blog、Twitter の影響は徐々に高まっている。特に Twitter の影響は徐々に上昇しており、先程の D の考察から潜在層を引き込むことになったのが長期的に人気となった一因であるだろうと考察できる。今回用いた GoogleTrends の数理モデルによる解析では、Blog と Twitter の影響は話題のピークを過ぎた period4、period5 で大きくなっている。この点はこれまでのヒット現象の数理モデルによる解析では出てこない新しい知見であるといえる。

5.2 ハンドスピナー

従来のヒット現象のモデルは常に D と P の値が高いことと、本研究で提案した数理モデルの P が period2 以降に高い値をとっていることから潜在層の人々に一気に広まり、顕在層へ移ったことから話題が絶え間なく続いたと考察できる。また、潜在層の間でも話題性が徐々に強まっていることがわかる。また、新しい数理モデルで D が period3 だけ低く、それに対して Blog の影響が period3 だけ大きいので、これまでのヒット現象の数理モデルではこの 2 つが区別できずに D に含まれていたと考えられる。

話題の皮切りとなった Yahoo ニュースなどから period1 でネットニュースの影響が高まっており、それを裏付けることができる。また、Twitter の影響も特に period1 に高まっており、それが徐々に低くなっていることからハンドスピナーのブームはネットニュースと Twitter が大きな要因になっていると言えるだろう。Tv の影響は period2、Blog の影響は period3 に高まっていることから拡散する過程でそれらの影響力が助けとなっていると考察する。

5.3 R_factor による精度比較

計算の R_factor は低い数値が出たケースほど精度よく GoogleTrends を再現していることを意味する。関心度の計算に Blog や Twitter の影響を入れた場合、入れない場合とで関心度の 4 つのパターンによる比較より、平均値と中央値において Blog と Twitter の影響を加えた方が精度が良くなった。特に Twitter を加えると精度が向上しやすい。従来の数理モデルと今回提案した数理モデルで比較すると、Blog の方がわずかに良くなった。Twitter の精度が悪いのはリツイートによって書き込み件数の振れ幅が大きく、リツイートを除いたオーガニックツイートとメンションに絞ることで精度が比較的よくなると考察する。以上から、人々の検索行動に Blog や Twitter は明らかに影響を与えていると考えられる。

6 結論

ヒット現象の数理モデルにならう形で、GoogleTrends の関心度を用いた Blog、Twitter の影響力を加えた検索の数理モデルを構築した。

その数理モデルを用いて分析を行った結果以下のことが分かった。

1. ブームの始まりにおける検索行動には Tv、ネットニュースの影響を受ける傾向にある。
2. Blog や Twitter の影響はブームが長続きするように働く。

また、R_factor による精度比較の結果より Blog、Twitter を加えた方が精度が良くなることが分かった。従来のモデルと比較して、Blog との差は 1 %以内だったが、Twitter より精度が良いことを示した。

参考文献

[1] Akira Ishii, Hisashi Arakaki, Naoya Matsuda, Sanae Umemura, Tamiko Urushidani, Naoya Yamagata and Narihiko Yoshida ."The 'hit' phenomenon: a mathematical model of human dynamics interactions as a stochastic process". New Journal of Physics 14(2012)

[2] Ishii A, Ota S, Koguchi H and Uchiyama K, the proceedings of the 2013 International Conference on Biometrics and Kansei Engineer- ing(ICBAKE2013) 143-147 DOI 10.1109/978-0-7695-5019-0/13 (2013)

[3] 大知正直・長濱憲・榊剛史・森純一郎・坂田一郎『口コミ指数による事例類型化に基づく複数メディアのヒット前の露出を先行指標とした情報拡散過程の分析』広報研究第 11 号 pp.35-50(2016)

[4] 石井晃・吉田就彦・新垣久史・山崎富美『ヒット現象の数理モデルとマーケティングサイエンス』 鳥取大学工学部研究報告第 37 号 pp107-113 鳥取大学 (2007)

[5] 漆谷たみこ『GRP を用いたヒット現象の数理モデルのパラメータ考察』（鳥取大学応用数理工学科　卒業論文）2010

[6] 福井佳奈・石井晃『ヒット現象の数理モデルを用いた「ピコ太郎」流行の解析』第 1 回計算社会科学ワークショップ (2017)