# 確率的ランキング
## ―流行度の順位付けとロングテール分析―

服部　哲弥

ウェブを電子的な店舗とするインターネット小売業などで，多数の商品などの流行度を反映するランキング（順位）を表示することが見られる．流行度を反映する，数学的にもっとも簡単な順位付けの数理モデルを考えると，モデルの単純さに比べて驚くほど，実際のランキングの時間変化の特徴を説明できる場合がある．この数理モデルの概要を紹介し，実際のデータを当てはめた結果を通して，商品の売上の分布のような，通常は社外秘に属する情報を，ランキングという公開されたデータだけから分析できる仕組みを説明する．

キーワード：確率順位付け模型，流体力学的極限，先頭に跳ぶ規則，ロングテール

### 1. Amazon のランキング

インターネット上の通販サイトの一つ Amazon.co.jp はインターネット書店の草分け的存在として出発した．サイト上で本を検索すると，その本についての紹介と注文用のボタン（リンク）を含むページが表示される．簡単のため以下では Amazon.co.jp の和書のページたちを「アマゾン書店」と呼ぶ．なお，本稿では，本の内容には一切興味がないので，「ページ」というときは実際の本を開く話ではなく，ウェブブラウザで表示されるウェブページを指す．アマゾン書店の一つのウェブ「ページ」が，町なかの書店店舗の陳列棚にある本の一点に相当する．

アマゾン書店の各書籍のページの中程やや下に，アマゾン書店が「ランキング」と呼ぶ，順位を表す数値がある．アマゾン書店は数百万ページ分の和書の表示事項を用意しているので，ほとんどの本の順位は数十万位から数百万位までの，各書籍それぞれの関係者以外にはあまり意味がないと思われる巨大な数値である．このランキングの数値は，アマゾン書店の説明（ヘルプのページ）や実際の観測によると，毎時 1 回変化する．すなわち，アマゾン書店のランキングは，時々刻々の時間変化をほぼリアルタイムで見ることのできる巨大な順位である．これは，インターネット時代以前は日常で見ることのなかった特徴である．

アマゾン書店はランキングの具体的な計算方法を公表していない．アマゾン書店の売上に基づいていて，売上が多いほど数値が小さく（順位が上であり），最近と過去の売れ行きを反映するという，誰もが当然視する内容の追認が説明にあるだけである．最近と過去の売れ行き，と説明しているが，観測によれば，最近の売れ行き，言い換えると，流行度を順位づけている，というのがおおかたの認識であろう．アマゾン書店の売上に基づくから，多数の読者によるランダムな注文が時々刻々の順位変化を定めることになる．

本稿はアマゾン書店のランキングのような**流行を反映する大規模な順位の時間変化**を興味の対象とする．

はっとり　てつや
慶應義塾大学 経済学部
〒 223–8521 横浜市港北区日吉 4–1–1

### 2. 確率順位付け模型

ランキング，すなわち流行を反映する大規模な順位の時間変化，のもっとも簡単なモデルとして，確率順位付け模型と呼ぶ多粒子系の確率過程を考える．本稿の背景にある研究の主題は，この模型について流体力学的極限に相当する解析を行うことである．得られた数学的結果を実際のアマゾン書店のランキング等に応用すると，例えば，アマゾン書店がロングテール・ビジネスと言えるか否かを分析できる．

以上について紹介するのが本稿の目的である．より詳しい内容は，2011 年 5 月に出版した拙著 [3] をご参照いただければ幸いである．原啓介さんによる 1 ページの書評 [2] も忙しい向きの参考になると思う．

#### 2.1 先頭に跳ぶ規則

「流行に応じた順位」の数学的にもっとも簡単な定義は，「商品が売れるたびにその商品を 1 位とすること」である．

直前に 1 位だった商品は 2 位に繰り下げ 2 位だった商品は 3 位などとすることによって順位の重複を解消

  

すれば，更新された重複や隙間のない順位を得る．このアルゴリズムは先頭に跳ぶ規則と呼ばれて，古くから研究されている [11].

流行とは文字どおり「最近もっとも売れている」ということであろうが，数学的に単純化・理想化して，まったく同時に 2 点以上の商品が売れる確率は 0 とすると，ある商品が売れたとき，その売れた時刻までの『十分短い時間』を考えれば，その時間で売れたその唯一の商品が一推しの流行となる．それまでどんなに人気がなくて順位が低い商品であっても，たまたま売れた瞬間に流行度 1 位とすることになる．

一方，購入者が不特定多数である状況を考えると，順位が 1 位に跳ぶ時刻の，すなわち商品が売れる時刻の，もっとも簡単なモデルは，各商品毎に独立に，ポワッソン確率過程に従うとすることである．このとき，同時に 2 点以上の商品が売れる確率は 0 となるので，数学的に矛盾のない定義になる．

1 つの商品が単位時間当たり 1 位に跳ぶ回数の期待値をポワッソン過程の用語で強度と呼ぶが，ここではわかりやすくジャンプ率と呼ぶ．簡単のために強度を定数として説明すると，ポワッソン過程とは，ジャンプ率が $w$ の商品が時刻 $s$ 以降時刻 $t$ までに $k$ 回 1 位に跳ぶ確率 $\mathrm{P}[\,\{k\}\,]$ が平均 $a=(t-s)\,w$ のポワッソン分布

$$\mathrm{P}[\,\{k\}\,]=e^{-a}\,\frac{e^{a\,k}}{k!}$$

になり，異なる時間区間の跳ぶ回数は独立な確率変数である確率過程をいう．

以上で「流行に応じた順位」の数学的にもっとも簡単な定義がすんだ．これを**確率順位付け模型**と呼ぶ [4–10]．なお，数理モデルの話をする間は，商品と呼ばず，味気なく粒子と呼ぶ．

図 1 は確率順位付け模型の粒子の動きの一例である．左端を 1 位，右端を最下位と対応させ，個々の粒子の名前（本のタイトル）を丸の中に書くことで図示した．初期状態から粒子 1，2，1，3 がこの順に 1 位に跳んだ例である．時間経過の後の並び方が最後に売れた順になることもわかる．少し前は「積ん読」，もう少し最近では「超整理法」として知られる原理とも同じである．このよく知られた原理を，アマゾン書店のランキングのような，流行度の順位付けのもっとも簡単な数理模型として採用しようということである．

本稿では立ち入らないが，参考までに，確率微分方程式を用いて確率順位付け模型の定義を書くと，ジャンプ率が時刻や順位の関数になる場合も定義を拡張できる．$N$ 個の粒子の系で，粒子 $i$ の時刻 $t$ における位置（順位）を $X_i(t)$，ジャンプ率を $w_j(X_i(t-),t)$，と置く．独立な $N$ 個の強度が 1 の一様なポワッソン過程を $\nu_j$, $j=1,2,\ldots,N$, また，事象 $A$ が成り立つ試行に対して 1，そうでないとき 0 となる確率変数を $\mathbf{1}_A$，と書くと，確率順位付け模型の時間発展は

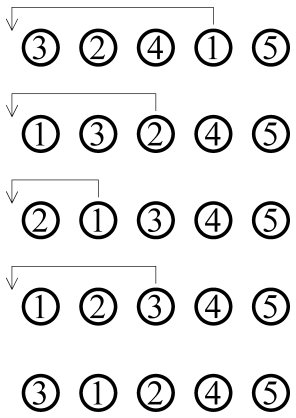


**図 1**　確率順位付け模型の粒子の動きの例

$$\begin{aligned}X_i(t)=x_i+\sum_{j=1}^{N}\int_{\substack{\xi\in[0,\infty)\\ s\in(0,t]}}&\mathbf{1}_{X_i(s-)<X_j(s-)}\\ &\times\mathbf{1}_{\xi<w_j(X_j(s-),s))}\,\nu_j(d\xi ds)\\ +\int_{\substack{\xi\in[0,\infty)\\ s\in(0,t]}}&(1-X_i(s-))\,\mathbf{1}_{\xi<w_i(X_i(s-),s))}\,\nu_i(d\xi ds),\\ i=1,2,\ldots,N,\ t\geqq 0,&\end{aligned}\tag{1}$$

で定義される [9]．右辺第 1 項 $x_i$ は粒子 $i$ の初期位置，第 2 項は粒子 $i$ より下位にいた粒子が 1 位に跳んだため粒子 $i$ の順位が下がること，第 3 項は粒子 $i$ が頻度 $w_i$ に応じてランダムに先頭 $X_i^{(N)}(t)=1$ に跳ぶこと，をそれぞれ表す．ジャンプ率の時刻依存性は例えば購買行動の昼夜差，位置依存性は例えばランキング上位にいることによる宣伝効果，をそれぞれこの数理モデルで扱えることを意味する．本稿では（1）については割愛する．

### 2.2　ランキングの時間変化―理論

先頭に跳ぶ規則は古くから研究されていたが，1 つの粒子が先頭に跳ばない時間に順位をどのように下げていくか，といった，時間変化を調べる視点は先行研究の関心の中心ではなかったようだ．ウェブ時代以前には巨大なランキングの時間変化が目に触れる機会がなかったため，順位低下の様子を観測する応用上の機会が乏しかったことも理由だろう．

ビッグヒット商品は順位が下がり始めてもすぐに注

文が入って 1 位に跳ぶことになるので，商品が売れない時間の順位低下を観測するのは難しい．

これに対して，ロングテールとも呼ばれる「売れないその他大勢」の商品は，言い換えると大多数の普通の商品は，売れない時間が長く，その順位は一般的な関心の対象にならない．まさにその時間依存性がここでの興味の対象となる．

時刻 $t$ までに 1 位に跳んだ粒子は一度も跳んでいない粒子の左側に位置する（たとえば図 1 参照）ので，特に，両者を分ける仕切りの位置が存在する．これを $X_C(t)$ と置く．時刻 0 に 1 位にいた粒子を $j$ とすると，すなわち $x_j = 1$ とすると，$j$ が 1 位に跳ぶまでは $X_C(t) = X_j(t)$ が成り立つ．ジャンプ率 $w_i$ たちが定数のとき，次の命題が成り立つ．

**命題 [5].** $N$ が大きいとき，$\frac{1}{N}(X_C(t) - 1)$ は $1 - \frac{1}{N}\sum_{i=1}^{N} e^{-w_i t}$ に近い． ◊

数学的には，両者の差が $N \to \infty$ の極限で 0 に確率 1 で収束するという命題である．言い換えると初期時刻に 1 位の粒子の位置は，次に 1 位に跳ぶまでの間は

$$X_j(t) \sim 1 + \sum_{i=1}^{N}(1 - e^{-w_i t}) \qquad (2)$$

と（命題の意味で）近似できる．左辺は確率変数だが右辺は決定論的であることに注意．この命題は，確率変数が決定論的な数に近づくという，粒子数についての**大数の法則**である．

ジャンプ率が時間の関数 $w_i(t)$ の場合でも，命題や（2）において右辺指数関数の肩を

$$w_i t \ \mapsto \ \int_0^t w_i(s)\, ds$$

と置き換えれば命題は成り立つ [8]．あらわに決定論的な公式が得られる理由は，（2）が成り立つ数学的な仕組みが独立確率変数の大数の法則だからである．ジャンプ率が位置の関数の場合は従属確率変数なので数学的な理屈はたいへん複雑になるが，この場合も大数の法則が成立することも最近わかった [9]．

数学的にも応用上も興味があるのは，ジャンプ率が，すなわち単位時間当たりの購入頻度が，商品によって異なる場合である．アマゾン書店の本は，ビッグヒットからロングテールとも呼ばれる「売れないその他大勢」の商品まで，平均的な売れ行きが単一ではなく分布する．ランキングの時間変化だけからその分布を推測できるか，という問題に肯定的に答えることができる．

値 $c$ に集中した単位分布（度数分布において，$c$ を含む区画に 1 単位の升を描くことの数学的理想化）を $\delta_c$ と書くと，ジャンプ率 $w_i$, $i = 1, 2, \ldots, N$，の分布は $\lambda_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{w_i}$ と書ける．この記号を用いると（2）の右辺の和は，$N\int_0^\infty (1 - e^{-wt})\, \lambda_N(dw)$ と書ける．

応用上は，$\lambda_N$ は，例えばアマゾン書店が並べている本の平均的な時間当たりの売上の分布である．これが，$N$ を大きくした極限である分布 $\lambda$ に近づく（数学的には，弱収束する）ならば，（2）はさらに

$$X_j(t) \sim N\int_0^\infty (1 - e^{-wt})\, \lambda(dw) \qquad (3)$$

と近似できる（重要ではない 1 は省略した）．すなわち，商品の**ランキングの時間変化は，商品の売上の分布のラプラス変換**として売上の情報を与える．特に右辺が $j$ と無関係なことに注意をお願いしたい．順位が下がる間の動きは，**どの商品のランキングの時間変化も同一**である．売れる商品と売れない商品のランキングの時間変化の違いは，（3）という共通の順位低下の流れからの離脱が早い（すぐ売れる）か，なかなか売れないか，の違いである．

### 2.3 ランキングの時間変化—データ

ここまで，流行度を反映する数学的にもっとも単純なモデルを紹介してきた．単純な数理モデルなので，1 冊売れただけで 1 位という顕著な特徴が，例えばアマゾン書店のランキングの振る舞いの良い近似になる保証はない．

『売れない専門書が一冊売れただけで一位になり，その後他の本が売れて順位を追い越すまでしばらく上位を占め続けるならば，ランキングの意味をなさない』という疑問が当然生じる．

実際のアマゾン書店のある本のランキングの時間変化のデータをグラフにしたのが図 2 である．横軸は時刻を表し，全体で約一年の長さである．縦軸はランキングを表し，数字の小さいほうが下というグラフの通常の描き方に従って図の上のほうが低順位である．縦軸の一番上の端が約 80 万位である．データ点の濃淡は，初期のデータ収集が著者の人力によっていたため，多忙で記録できないことを反映する．前小節で指摘し

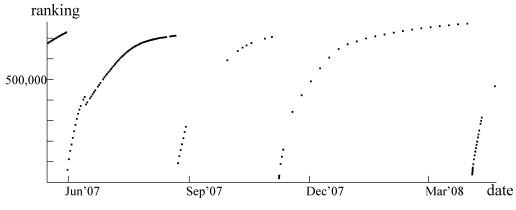


**図 2** アマゾン書店のランキングの時間変化の例

  

たように，図 2 は 1 点の本を**任意**に選んで順位変化を追いかければよい．帆船の帆のような湾曲した曲線の形はどの本を選んだかによらない（やや皮肉なことだが，人気の無い本を選んだほうが軌道の形状が長時間観測できるので研究上は望ましい）．

一目瞭然，順位が悪化する（数値が増える）ときは，帆船の帆のような，ほぼなめらかな，上に凸な増加曲線に沿って変化し，順位が改善するときは一気に横軸付近まで跳ぶ．順位の一気の改善がアマゾン書店におけるその本の注文行動に対応することは，人気のない専門書を注文して 2 時間ほど順位の変化に注目すればすぐにわかる．こうして，確率順位付け模型という，数学的にもっとも単純な順位付けのモデルの特徴が，実際のアマゾン書店のランキングの観測事実として実在することがわかる．

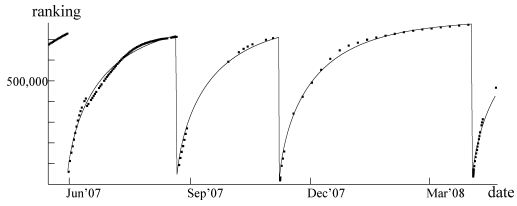


**図 3**　理論曲線のあてはめ

確率順位付け模型という数学的な単純さを追求したモデルが意外に複雑な現実のデータの特徴を説明していることがわかったので，もう一歩踏み込んでデータを理論式（3）に統計的に当てはめてみる．

式（3）によると，確率順位付け模型に基づく順位の時間変化（理論曲線）はジャンプ率の分布 $\lambda$ がわかれば求まる．アマゾン書店でいえば，書店が用意する本たちそれぞれの平均的な売上を本について集計した売上分布が $\lambda$ である．しかし，商品の売上分布のような情報は，店の経営陣は直接的に把握できるが，社外秘に属するので著者にはわからない．そこで，数学の道筋とは逆に，図 2 の実測データを（3）に統計的に当てはめることで，アマゾン書店の売上分布 $\lambda$ を推測することを考える．

数学的にもっとも単純なモデルを考えたので，$\lambda$ もできるだけ単純な分布を選ぶ．アマゾン書店がロングテールビジネスの草分け的存在として注目されることがあることを考えて，$\lambda$ として（一般化）パレート分布（離散版 $\lambda_N$ として一般化ジップの法則）を選ぶ：

$$\lambda([w,\infty)) = \left(\frac{a}{w}\right)^b, \quad w \geqq a. \tag{4}$$

$a$ と $b$ は正定数である．理論分布（4）を（3）に代入すると

$$X_j(t) \sim N - Nb(at)^b\Gamma(-b, at)) \tag{5}$$

と，不完全ガンマ関数 $\Gamma(z,p) = \int_p^\infty e^{-x}x^{z-1}\,dx$ を用いて理論曲線が求まる．図 2 のデータを用いてパラメータ $N$, $a$, $b$ を求めることで，

$$(N,\ a,\ b) = (8\times10^5,\ 5\times10^{-4},\ 0.77) \tag{6}$$

を得た．$N$ は書籍点数，$a$ は時間の逆数，$b$ は次元を持たない指数である．これを用いて理論曲線を図 2 の実際のデータに重ねたのが図 3 である．思い切り単純な数理モデルにしては，定量的にも現実のランキングのデータをよく説明する，と考える．

## 3. アマゾンはロングテールに非ず

世にあるほとんどの本はめったに売れない．数百万点におよぶ和書のうち，町なかの普通の規模以下の書店が扱うのは一握りのビッグヒットである．

他方，ビッグヒットを除く個々の本はめったに売れなくても，そのような本はきわめて多数あるので，合計すれば経営上無視できない売上をもたらすのではないか，というのがロングテールビジネスの可能性である．ウェブ小売業以前は商品陳列のためのコストが高かったので，どのみち多数の商品を置くことはできず，ビッグヒット依存型の商売しかありえなかったのに対して，アマゾン書店を含むウェブ小売業は，ウェブページによる「商品陳列」を行うことから，商品一点ごとのコストが大幅に下がり，ロングテールビジネスの可能性が現実的な検討課題になった．

アマゾン書店は，ロングテールビジネスの草分け的存在として注目されたことがある [1]．多数の本のページをもつアマゾンは，めったに売れない多数の本の売り上げが無視できないかもしれない．この可能性を検証するためには，アマゾン書店における本の売上の分布を知る必要がある．商品の売上分布のような情報は，店の経営陣は直接把握できるが，社外秘に属する．ところが，ランキングという公開情報だけを用いることで得た（4）と（6）は，まさにアマゾン書店の売上分布（の近似）である．つまり，アマゾン書店がロングテールビジネスであるか否かを部外者でありながら分析できる．

図 4 は横軸によく売れる順に商品を並べ，縦軸にそれぞれの商品の売上をとったときの売上曲線の概念図である．縦軸に平行な線と売上曲線が囲む面積が，対応

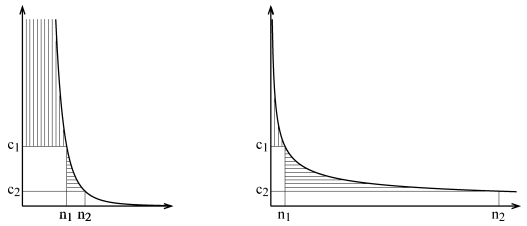


**図 4**　ロングテールの売上への寄与と指数 $b$

する商品たちからの売上への寄与を表す．ロングテールの寄与は，図の左端付近にある一握りのビッグヒットを除いた，図の右のほうの面積となる．パレート分布族（4）の場合は，全体の売上の中でロングテールが占める割合を決めるのは指数 $b$ である．$b$ が小さいとき図 4 の左図のように，裾野に比べてビッグヒットの寄与が圧倒的であり，$b$ が大きいときは右図のようにロングテールが無視できない．$N$ が大きいとき，ロングテールの売上への寄与が全売上の中で無視できるかどうかは $b$ が 1 より大きいか小さいかが判定基準となることがわかる．データを当てはめた結果（6）から $b < 1$ とわかったので，アマゾン書店の場合はロングテールの売上は無視できる．**アマゾンはロングテールに非ず**ということである．

アマゾンのランキングから $b$ を求める先行研究に $b > 1$ と結論しているものが複数見られたが，本稿で紹介した確率過程の考察を経ておらず，ランキングの時間変化についての粗雑な解釈に基づく誤った結論である．アマゾン書店は，その膨大なカタログのページ数にもかかわらず，売上の事実上すべてを一握りのビッグヒットが支えている．

以上は理論の枠組みのうちでもっとも単純な部分の紹介である．より立ち入った応用上の興味としては，例えば，ジャンプ率に時刻依存性を入れることで，活動の昼夜差をランキングの動きだけから分析することができる．直感的には，人々の購入活動が活発ならば 1 位になる商品が素早く入れ替わるので，購入されない商品の順位の下がり方は激しくなる．巨大掲示板 2ch.net のスレッド一覧のデータを詳細に分析したところ，図 5 のように，夜間真夜中までの活動が活発で，真夜中を過ぎて未明の時間帯は動きが鈍い，という結果を得た [8][3]．この傾向はアマゾン書店でも見ることができる．ネット活動が昼夜逆転しているといったことはなさそうである．

## 4. 流体力学的極限

本稿を終える前に，この数理モデルに対する数学的

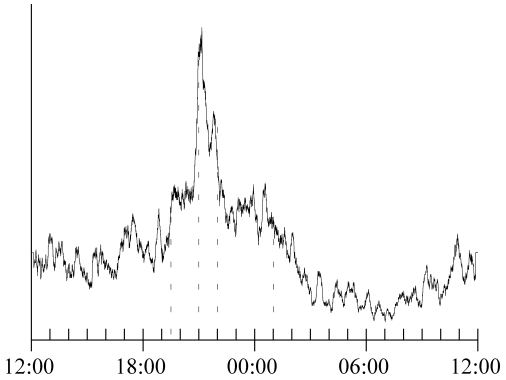


**図 5**　2ch.net のスレッド一覧に見る活動の昼夜差

な興味を少し紹介したい．確率的な順位付けの模型を考えたため順位の上位に売れない本が来ることもあれば，本来ならよく売れるはずの本が下位にいることもある．しかし，概してビッグヒットは一時的に順位を下げてもすぐ売れて上位に戻るし，売れない専門書は 1 冊売れて 1 位になっても次の幸運がなかなかないので，多くの時間は下位にとどまる．本の点数 $N$ が大きくなれば大数の法則によってこの傾向は安定することが期待できる．これを数学的にとらえるために粒子の位置とジャンプ率の結合経験分布

$$\mu_t^{(N)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{((X_i(t)-1)/N, w_i)}$$

を考えると，初期状態の収束 $\lim_{N\to\infty} \mu_0^{(N)} = \mu_0$ の下で，任意の時刻での収束 $\lim_{N\to\infty} \mu_t^{(N)} = \mu_t$ が成り立つ [5]．このことはジャンプ率が時空依存性をもつ一般の（1）で成り立ち [8][9]，さらに各点毎の収束だけでなく確率過程としての収束も成り立つ [8–10]．有限粒子系の分布 $\mu_t^{(N)}$ はランダム（分布値確率過程）だが，極限 $\mu_t$ は決定論的であって，その分布関数は，inviscid Burgers 型に似た，ある 1 階準線形偏微分方程式の解として特徴づけられる [6][8][9]．

ミクロな視点では多数の粒子がランダムに運動する系が，マクロな視点ではなめらかで決定論的な連続体の運動に見える．この意味で分子運動と流体の流れの二つの描像を結びつける**流体力学極限**に似る．数学的にはそのもっとも簡単で非自明な例題を発見したと位置づけられるだろう．

本稿で紹介した研究は，純粋に数学的なモデルの数学的な結論を当てはめることによって，ランキングという限られたデータだけからロングテール・ビジネスの成立不成立という経営上も興味深い情報を得る．も

  

ちろん「中の人」にとっては直接入手できる情報だが，通常は社外秘に属するであろう経営上の貴重な情報を，ランキングという公開された情報だけから合法的に分析できる．

また，経営側にとっても，実はロングテール部分の個々の商品の平均売上を正確に計測することはできないので，例えばリストラの際に扱う商品を大きく減らす必要が生じたときに切る商品の選択の合理的判断が難しい．これに対して，流体力学的極限の結果から，ロングテールビジネスが不成立でビッグヒット依存型の場合は，例えばある任意の時刻のランキングにおいて下位の商品を切るという単純な方法が経営上十分良いことがわかる．数学的な深い議論も現実の問題とかかわりがある．

## 参考文献


[1] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*, Hyperion Books, 2006.

[2] 原　啓介，書評「Amazon ランキングの謎を解く」，数理科学，**581**, 2011 年 11 月号，p. 61.

[3] 服部哲弥，『Amazon ランキングの謎を解く』，化学同人，2011. `http://web.econ.keio.ac.jp/staff/hattori/amazonj.htm`

[4] T. Hattori, *Stochastic ranking process and web ranking numbers*, in *Mathematical Quantum Field Theory and Renormalization Theory,* T. Hara, T. Matsui, F. Hiroshima, eds., Kyushu University web, `http://gcoe-mi.jp/publish_list/pub_inner/id:2`, Math-for-Industry Lecture Note Series, **30** (2011), 178–191.

[5] K. Hattori and T. Hattori, *Existence of an infinite particle limit of stochastic ranking process*, Stochastic Processes and their Applications, **119** (2009), 966–979.

[6] K. Hattori and T. Hattori, *Equation of motion for incompressible mixed fluid driven by evaporation and its application to online rankings*, Funkcialaj Ekvacioj, **52** (2009), 301–319.

[7] K. Hattori and T. Hattori, *Sales ranks, Burgers-like equations, and least-recently-used caching*, RIMS Kokyuroku Bessatsu, **B21** (2010), 149–162.

[8] Y. Hariya, K. Hattori, T. Hattori, Y. Nagahata, Y. Takeshima and T. Kobayashi, *Stochastic ranking process with time dependent intensities*, Tohoku Mathematical Journal, **63–1** (2011), 77–111.

[9] T. Hattori and S. Kusuoka,*Stochastic ranking process with space-time dependent intensities*, preprint, 2011.

[10] Y. Nagahata, *Tagged particle dynamics in stochastic ranking process*, preprint, 2010.

[11] M. L. Tsetlin, *Finite automata and models of simple forms of behaviour*, Russian Mathematical Surveys, **18** (1963), 1–27.