Amazon.co.jpのランキングのモデルとロングテールの分析

服部哲弥　（東北大学・理）
服部久美子　（首都大学東京・数学）

2008.09.02 研究集会「紀要の電子化と周辺の話題」（京大数理研）

START

1．Amazon.co.jpのランキング

Amazon.co.jp

本のページ中程やや下

Amazon.co.jp **ランキング**

「Amazonの謎順位。」

‘Internet retailers are extremely hesitant about releasing specific sales data’

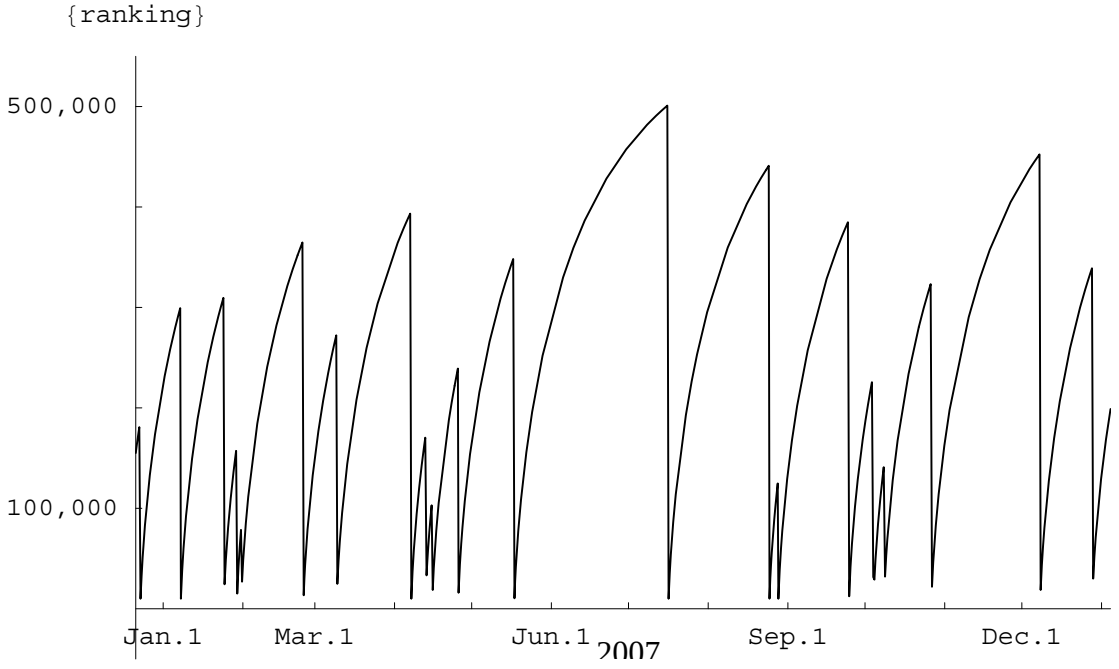
ランキングの時間変化

- **本を書くと，自分の本の順位が気になる．**

Amazon.co.jp ランキング：本で373,406位

(1時間後) Amazon.co.jp ランキング：本で373,977位

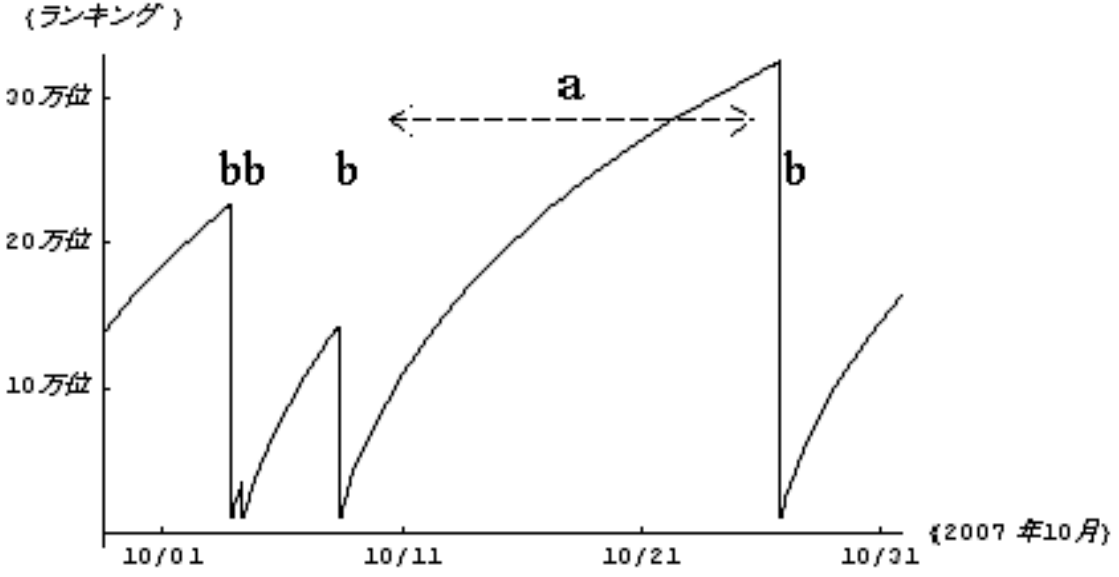
(2 時間後) Amazon.co.jp ランキング：本で374,693位



- **Amazon.co.jp のランキング（順位）変化はワイルド，ランダム**

ランキングの時間変化のモデル化

- できるだけ単純化したモデルで本質を理解したい



- Stochastic ranking process
  - a. 売れない間は他の本が売れて追い越すたびに順位が下がる
  - b. 売れると即座に1位

Stochastic ranking process

a. **売れない間は他の本が売れて追い越すたびに順位が下がる**

b. **売れると即座に１位**

- **これだけで「Amazon 謎順位」が説明できる**
- **Amazon.co.jp ランキングの時間変化は定量的な情報を持つ**

今日の話： Stochastic ranking process**は，特にロングテール（極めて多数のそれぞれは少ししか売れない本たち）について，**

- **定量的分析に有効（確率モデルの多粒子極限→偏微分方程式）**
  - ○ Pareto **指数の決定（「Amazon.co.jp はロングテールビジネスか？」に答えられる）**
  - ○ **ランキング下位の売り上げへの寄与の計算が可能（マネージメント；短時間で決断したときの損失評価）**
- **単純だが自然な順位付け（個別のランダムと全体の決定論）**

＊．目次

2. Stochastic ranking process

N 個の粒子の列の順序（順位，ランキング）変化の確率モデル

- **定数：**

$x_{1,0}^{(N)}, \cdots, x_{N,0}^{(N)}$; $1, 2, \cdots, N$ **の並べ替え　（順位の初期値）**

$w_1^{(N)}, \cdots, w_N^{(N)}$; **非負　（1 位への** jump **率＝本の注文頻度，…）**

- **確率変数：**

$X_1^{(N)}(t), \cdots, X_N^{(N)}(t)$　　**（時刻** t **での各粒子の順位）**

[規則 0**]**　$X_i^{(N)}(0) = x_{i,0}^{(N)}$ $(\forall i)$

$\tau_{i,j}^{(N)}$, $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots$;

（粒子 i **が** j **回目に 1 位に** jump **する時刻）**

j **について増加**

$\tau_{i,j+1}^{(N)} - \tau_{i,j}^{(N)}$, $j = 0, 1, 2, \cdots$ $(\tau_{i,0}^{(N)} = 0)$ **は** i, j **について独立，** j **について同分布**

で指数分布　$\mathrm{P}[\, \tau_{i,1}^{(N)} \leqq t \,] = 1 - e^{-w_i^{(N)} t}$

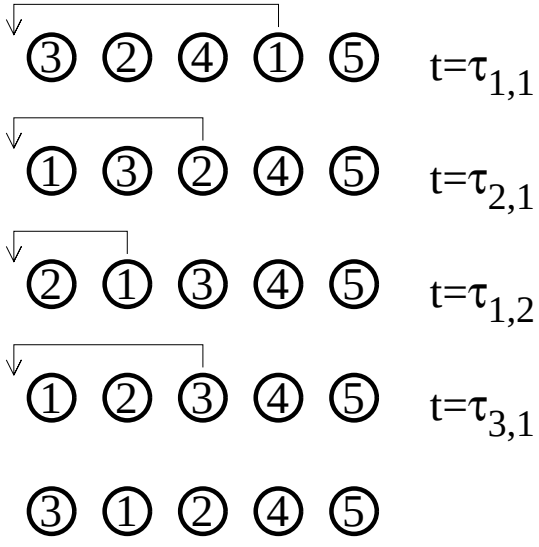
時間発展の規則

[規則 1] $X_i^{(N)}(\tau_{i,j}^{(N)}) = 1 \ (\forall i, j)$

[規則 2] $X_i^{(N)}(\tau_{i',j'}^{(N)}) = X_i^{(N)}(\tau_{i',j'}^{(N)} - 0) + 1 \ (\forall i, i', j')$

（各粒子は自分より下位の粒子がjump**するごとに順位を**1**下げる**

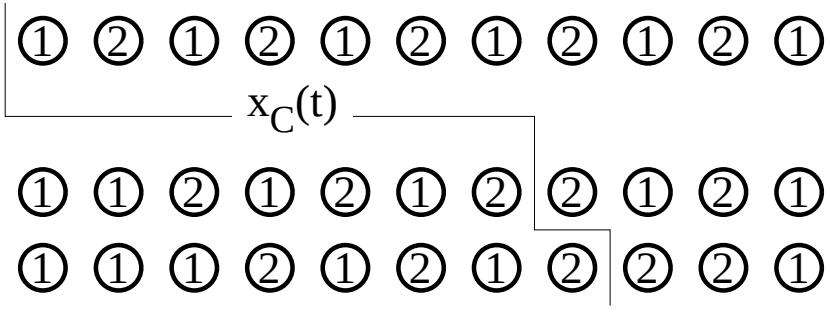
＝jump**以外では列の相対順序保存）**



$\tau_{1,1} < \tau_{2,1} < \tau_{1,2} < \tau_{3,1} < \cdots$ **なるサンプル**

$x_C(t)$：jump済み粒子と未jump粒子の境界

$x_C(t)$：jump済み粒子と未jump粒子の境界



$$x_C^{(N)}(t) = 1 + \sum_{i=1}^{N} \chi_{\tau_i^{(N)} \leqq t}$$

（1位になった時刻を$t = 0$に取り直すと，）

$X_C^{(N)}(t) =$ 1位になった粒子のその後の軌道

$x_C(t)$の大数の法則

ランダムなモデル $\Rightarrow$ Nが大きいとき決定論的な運動（大数の法則）

Jump率の分布 $\lambda^{(N)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{w_i^{(N)}}$ が $N \to \infty$ で λ に弱収束するならばス

ケールした軌道 $y_C^{(N)}(t) = \frac{1}{N}(x_C^{(N)}(t) - 1) = \frac{1}{N}\sum_{i=1}^{N} \chi_{\tau_i^{(N)} \leqq t}$ は

$y_C(t) = 1 - \int_0^{\infty} e^{-wt}\lambda(dw)$ に弱収束する.

☆　$y_C(t)$ は実際に観測される！

データへの当てはめにはjump率（＝販売頻度）の分布 λ が必要（後述）

経験分布の収束

仮定．　初期配位 $y_{i,0}^{(N)} = \frac{1}{N}(x_{i,0}^{(N)} - 1)$ **の分布が** $N \to \infty$ **で収束：**

$$\mu_{y,0}^{(N)}(dw\,dy) = \frac{1}{N}\sum_i \delta_{w_i^{(N)}}(dw) \otimes \delta_{y_{i,0}^{(N)}}(dy) \to \mu_{y,0}(dw) \times dy \ (N \to \infty)$$

定理：　Jump 率と相対順位 $Y_i^{(N)} = \frac{1}{N}(X_i^{(N)} - 1)$ **の結合経験分布（分布値確率変数列）** $\mu_{y,t}^{(N)} := \frac{1}{N}\sum_i \delta_{w_i^{(N)}} \otimes \delta_{Y_i^{(N)}(t)}$ **は** $N \to \infty$ **で（非ランダムな）結合分布** $\mu_{y,t}(dw) \times dy$ **に確率収束する**

☆　**極限** $\mu_{y,t}(dw)$ **はあらわに分かる**　←　**偏微分方程式の解**

極限を記述する偏微分方程式

$$\frac{\partial U_i}{\partial t}(y,t)+\sum_j f_j U_j(y,t)\frac{\partial U_i}{\partial y}(y,t)=-f_i U_i(y,t) \qquad (y,t)\in[0,1)\times[0,\infty)$$

1次元非圧縮性混合流体の蒸発による運動，Burgers型方程式
定数： $f_i \geqq 0$ **（第i種流体の蒸発率）**
未知関数： $U_i(y,t)$ **（時刻tにyより右にある第i種流体の量）**
初期値： $U_i(y,0)\geqq 0$, smooth, $\searrow$, $\sum_j f_j U_j(0,0)<\infty$,

$\sum_j U_j(y,0)=1-y$ **（Burgers型システムで衝撃波の無い初期値）**

境界条件： $U_i(0,t)=U_i(0,0),\ t\geqq 0$ **（定常）**

☆ **解が唯一存在（特性曲線であらわに解ける非線形システムの例題）**

特性曲線 $\frac{dy_B}{dt}(t)=v(y_B(t),t);\ v(y,t)=\sum_j f_j U_j(y,t),\ y_B(0)=y_0$

以上はjump率が離散分布の場合： $\mu_{y,t}(\{f_i\})=-\frac{\partial U_i}{\partial y}(y,t)$

極限を記述する偏微分方程式（一般のjump率分布）

$$\frac{\partial U}{\partial t}(dw; y, t) + \int w' U(dw'; y, t) \frac{\partial U}{\partial y}(dw; y, t) = -wU(dw; y, t),$$

$(y, t) \in [0, 1) \times [0, \infty)$

初期値： $U(dw; y, 0) \geqq 0$ smooth, **非増加** in y, $\int wU(dw; 0, 0) < \infty$, $U(\mathbb{R}_+; y, 0) = 1 - y$ **（Burgers型のシステムで衝撃波の起きない初期値）**

境界条件： $U(dw; 0, t) = U(dw; 0, 0),\ t \geqq 0$ **（定常）**

☆ **解が唯一存在（特性曲線であらわに解ける非線形システムの例題）**

$$\mu_{y,t}(dw) = -\frac{\partial U}{\partial y}(dw; y, t)$$

定理の意味

- 順位の先頭付近はjump率の高い粒子が多くtailは低い粒子が多い

- ランダムな現象がNが大きいとき決定論的運動に近い

（Amazon.co.jpの本は百万冊の程度だから十分適用可能）

確率モデル：　売れる本・売れない本の順位分布を追うのに全ての注文記録が必要

↓

PDEの解：　初期配置があれば個別の注文記録は不要

- 従属確率変数の大数の法則，空間分布の経験分布の収束

（右側の粒子が飛んだときだけ位置がずれる）

３．データへの当てはめとロングテール分析

粒子の軌道（本の順位），最後に売れた時刻を$t = 0$と取り直す

$$x_C(t) \simeq N y_C(t) = N - N \int_0^\infty e^{-wt} \lambda(dw)$$

実際のデータに当てはめるためには，Jump率の分布 λが必要

社会学や経済学ではPareto分布が使われることが多い

Jump率の分布がPareto分布のとき

Pareto分布： $w_i = a\left(\frac{N}{i}\right)^{1/b}$, $i = 1, \cdots, N$, $a, b > 0$は定数

例：w_iはi番目の金持ちの年収　　「80–20の法則」

a: 最低収入　（最高収入$= aN^{1/b}$)

指数b：　b小$\Leftrightarrow$不平等，　b大$\Leftrightarrow$平等

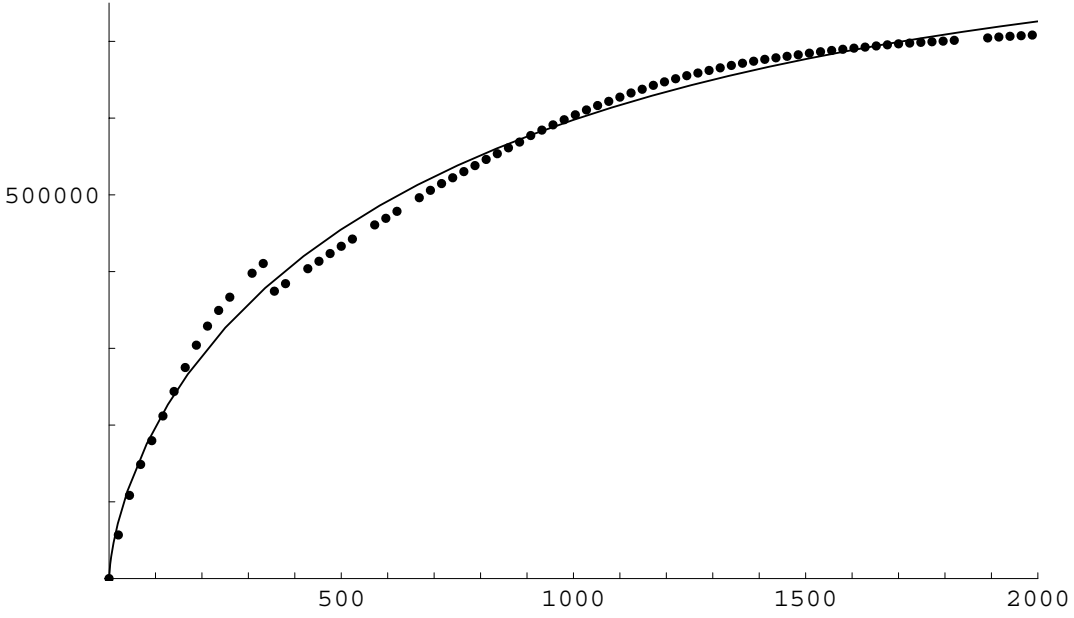
$x_C(t) \simeq N\, y_C(t) \simeq N(1 - b(at)^b \Gamma(-b, at))$

（Γ：不完全ガンマ関数）

N, a, bを与えれば決まる！

ランキングデータへの当てはめ

2000時間（80日）超のデータ（$n_d - 1 = 77$）



$N^* = 90$万

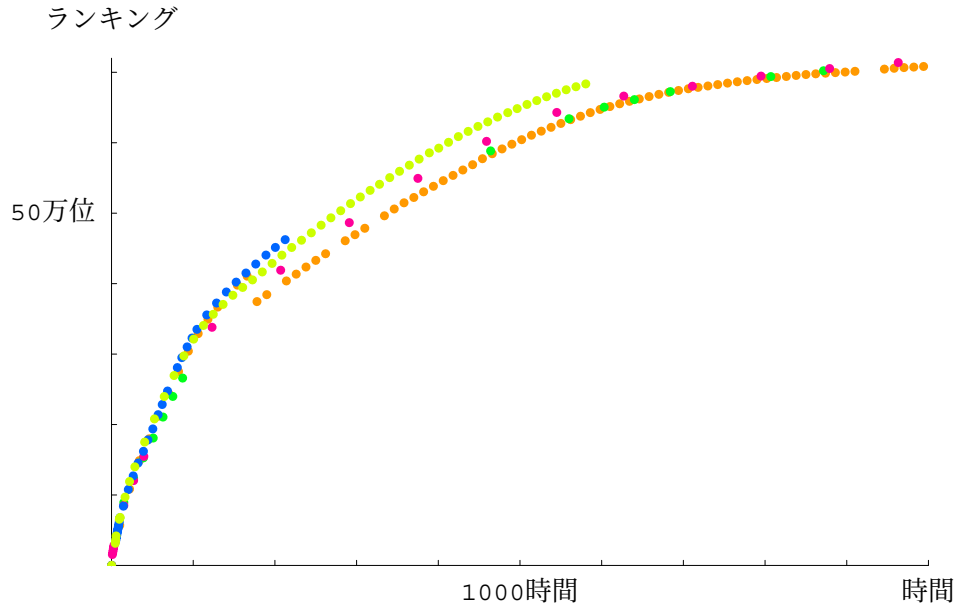
$a^* = 4 \times 10^{-4}$

$(1/a^* = 3.5$ヶ月$)$

$b^* = 0.6312$

$(\sqrt{\chi^2/n_d} = 1$万$)$

再現性



概ね良いが，長期に関してはトレンドが無視できないように見える．短期に関しては日変化（午前午後）の波も無視できないようだ.（今後の問題）

先行研究との比較

先行の経済学的研究
（選んだ本の期間あたりの平均販売量を多数の本にわたってとる方法）

Chevalier − Goolsbee　$b = 1.2$

Online bookstore が brick-and-mortar bookstore より価格弾力性が高く，CPI への影響大

Brynjolfsson − Hu − Smith　$b = 1.148$

J. A. Hausman (1997) の消費厚生 (consumer welfare) 評価方法：Long tail 領域の書籍の購入が可能になったことを値段が需要 0 の高値から下がったと計算（価格弾力性は input）

これらの b の値は我々の方法＋実測値とは矛盾

Online retail の経済効果は先行主張ほど強くない？

少なくとも Amazon.co.jp については，ロングテールの経済効果は小さい（後述）

テールからの売り上げへの寄与の評価

・ ここまで特性曲線 $y_C(t)$ の応用　－　$\mu(dw; y, t)$ の応用？
引き続き，λ はPareto分布，単価は等しいと仮定．$0 < r < 1$ に対して
定常に達した後のランキングの下位（上位rを除いた残り）からの総売上への寄与：

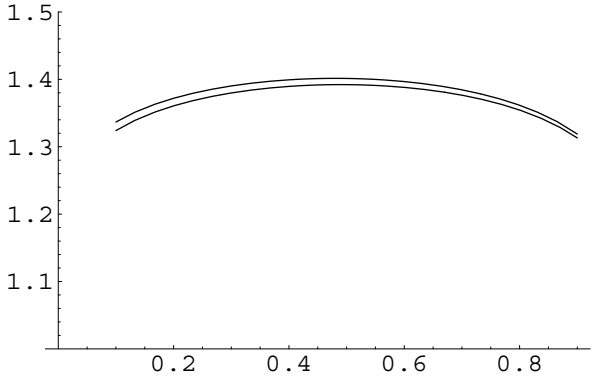
$$\tilde{S}(r,1) \simeq N \int_{(w,z)\in[0,\infty)\times[r,1)} w\mu_{z,t}(dw)\,dz$$
$$= Nab\,\Gamma(1-b, q(r))\,q(r)^{b-1}$$
$$;\ q(r) = a\,t_1(r),\ r = 1 - e^{-q(r)} + q(r)^b\,\Gamma(1-b, q(r))$$

cf.　真の売れ行き w_i を知っていてその順に並べた場合の下位 $1-r$ からの寄与

$$S(r,1) \simeq N\frac{ab}{b-1}(1 - r^{(b-1)/b})$$

右図：$\dfrac{\tilde{S}(r,1)}{S(r,1)}$ $(b = 1.15,\ 1.2)$

→　（ランキングの下位を切って捨てたときの損失が正確な売り上げ順位に従って切った場合の最大1.4倍程度）



指数bとロングテール

b小$\Leftrightarrow$不平等，　b大$\Leftrightarrow$平等＝ロングテール　　もう少し定量的に

$b > 1$: $S_{tot} = S(0,1) \simeq \dfrac{Nab}{b-1}$ **収束**

極端な例 $b = 2$: $\dfrac{S(0,0.2)}{S_{tot}} \simeq \sqrt{0.2} \simeq 0.447$

→　テールは売り上げの半分を占める（20–80の法則からの大きな逸脱）

現実的な数字　→　20–80の法則に近い：

$b = 1.2$ (Chevalier, Goolsbee) $\dfrac{S(0.2,1)}{S_{tot}} \simeq 0.235$,

$b = 1.15$ (Brynjolfsson, Hu, Smith) $\dfrac{S(0.2,1)}{S_{tot}} \simeq 0.189$

しかし，N自体が大きい状況なので無視はできない　→　ロングテールビジネス

$b < 1$: $S(r,1) \simeq \frac{Nab}{b-1}(1 - r^{(b-1)/b})$: $r = 0$で**発散**

ヒット商品が効く．**ロングテールビジネス不成立**

Amazon.co.jp の観測結果：$b^* = 0.6312 < 1$

Amazon.co.jpはロングテールビジネスか？

C. Anderson, 'The Long tail'

Amazon.co.jp: $b^* = 0.6312 < 1$

インターネットを生かしたロングテール型のリテールの草分けとして有名なAmazon書店は，ロングテールビジネスでは無い

Amazon はロングテールの「成功例」として記事や解説や論文に登場することで宣伝費無料の大規模宣伝を勝ち取り，実際にはハリー・ポッターの予約販売（旧態依然の大ヒットビジネス）で利益を上げている，と思われる．

傍証： Amazonが売上げ詳細を隠すと言われている…

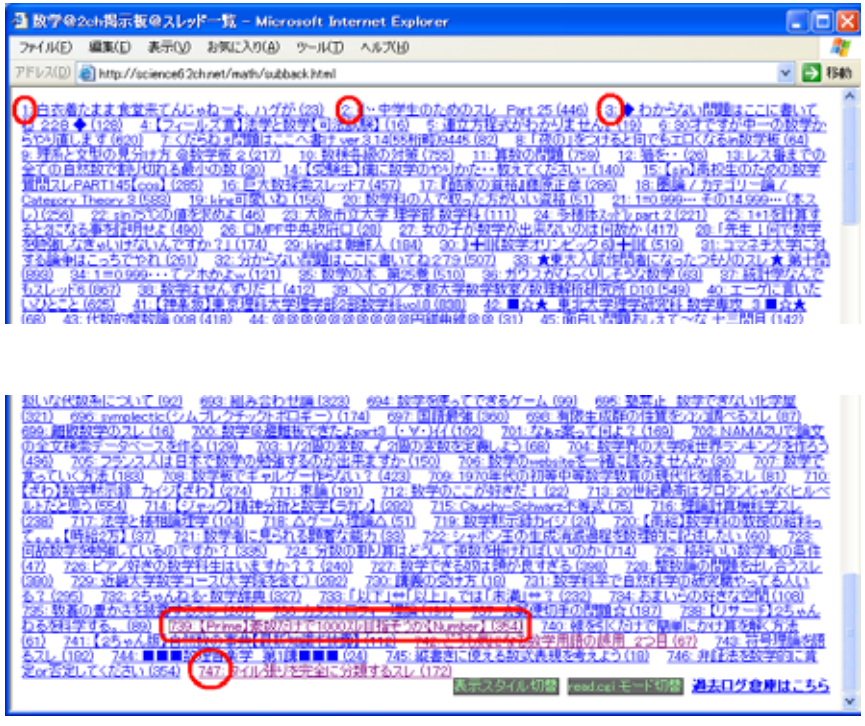
おまけ．2ch.net

２チャンネル (2ch.net)：
web **掲示板の巨大な集まり**

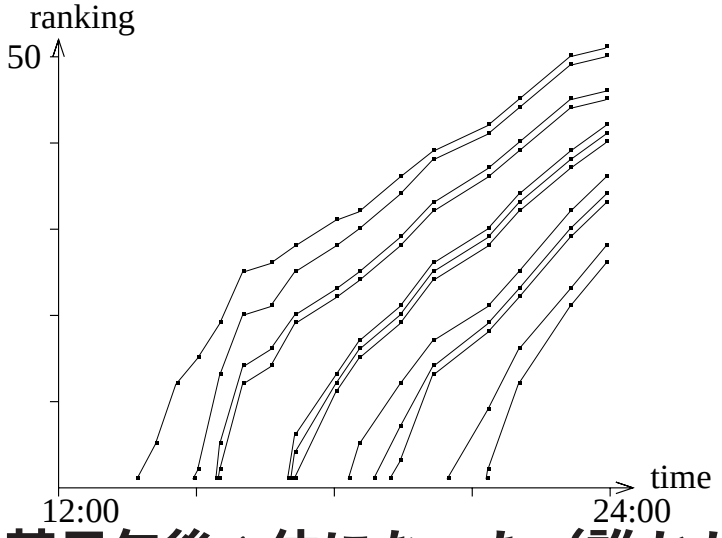
スレッド（ページ）一覧：
「書き込んだスレが１位」：

stochastic ranking process

と同じ

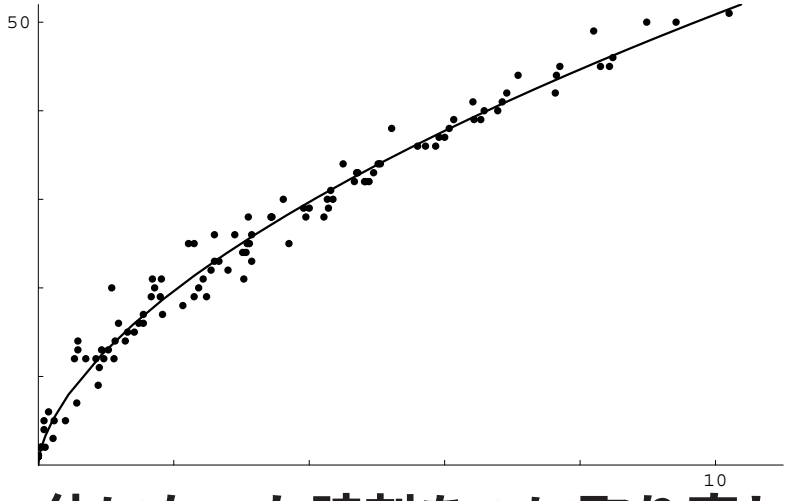


スレ一覧の順位変化



某日午後１位になった（誰かが書き込んだ）スレのその後の順位の時間変化

（複数回１位になったスレについては最後に１位になって以降の時間変化）



１位になった時刻を０に取り直して重ねた図

実線：モデルの無限粒子極限から得られる曲線

$N = 795,$

$(a^*, b^*) = (3.3 \times 10^{-4}, 0.62)$

$(1/a^* =$約４ヶ月$)$

☆　ブログの人気度ランキングにも応用可能

4．まとめ

stochastic ranking process：Amazon.co.jp ランキング （会社が公表しない，テール側の書籍にとって一見ワイルドな） 「謎の順位」を説明する簡単なモデル

揺らぎの大きい多種類のデータがあるとき，ラプラス逆変換によって，種類の分布 λ を揺らぎ無しに決める統計手法．

- 確率過程論 (stochastic ranking process)
- 偏微分方程式論（ランダム系から決定論へ）
- 数理統計学（ラプラス変換を用いたデータ解析）
- 計量経済学（online retail, long tail の分析）

- 事実の定量の重要性（一見複雑な現象が，単純なモデルで（半）定量的に説明できる．誤解や無用の複雑な話が出回る前に，何が数学的に単純かを公にする価値）
- 数学的モデルが刺激になって，新しいデータが発表されることを期待

応用の可能性

- 情報公開を渋る会社の経営状況を探る
- ロングテール型オンラインリテールの業績情報公開のための法的規制の方法（安価で単純な仕組みで，ロングテール構造が分かる）
- 大量の統合されずに蓄積されたデータのデータマイニング.
  - ○ 会社のwebや個人単位のブログのランキング
  - ○ 管理人が管理しきれない掲示板のactivity分析
  - ○ 緊急時の安否などの情報（フォーマットに従う余裕はない）

正確な情報や個別具体的な情報は探索の問題だが，統計的な情報（ベストヒットとロングテールの売り上げへの貢献の比，どの地域が情報弱者が多いかなど，ネットへのアクセス頻度分布）は，最小限の構造だけを仮定してすばやく測定したい

課題（例）

時間変化

- マクロな量（総売上，種類数，販売率頻度）のトレンド（新規出版年８万冊）や周期などは（数学的に複雑になるが）データから解析できるはず
- 個別の本のライフサイクル（栄枯盛衰）はノイズが大きいからデータから確実に導出できるとは思えない．
- ライフサイクルがあっても分布λが定常ということは可能．その場合，時間変化を考慮しない単純なstochastic ranking processがどれくらい良い近似か？上位と下位のランキングの時間発展のずれなどで検出可能か？

Stochastic ranking process からの示唆(?)

本題は以上でおしまい．以下は余談．

- Stochastic ranking process は応用上，long tail （売れ行きの小さいあまたある品々）の定量的分析の手段として有効．
- 判断の材料となる客観的な統計的手法を提供するものであって，社会政策や「売れる方法」を提案するものではありませんが…．
- 「めったに売れないあまたある専門書」の価値や「ほとんど引用されない数多くの学術論文」の価値を信じるからこの研究に深入りしたので，可能ならば何かを示唆したい…．

ランキングと文化

厳しい現状.

- **ランキングに敏感な時代の危機感**
  - ○ NHK クローズアップ現代（2008 年 6 月 4 日 No.2592）「ランキング依存が止まらない〜出版不況の裏側〜」
    - 売り上げランキングをもとに本を選ぶ人や下位の本を返品する書店の例
    - 出版点数の急増や出版社倒産などの問題

  （本当にリンクしているかは吟味が必要だが…）
  - ○ 'Citation Statistics', (2008.6) International Mathematical Union (IMU), ICIAM, IMS
    - 「単純・客観的」な citation data の科学研究評価における 'use and misuse'
  - ○ 出版点数や論文数の増加と構造特区でのタクシー増加
- **大ヒット依存ビジネスモデルからの脱却**
  - ○ 出版業界も研究者社会も 20–10 年前が良すぎたと考えるべき？
  - ○ C. Anderson (long tail)：大ヒット無き時代のポップス界

☆ 経済効果を考えれば，ランキング上位が重要．Long tail はヒットが出なくなったときに「堪え忍ぶ」話と見える．

統計量で測れるものと漏れるもの

厳しいから，守るための示唆や提案が必要．

- 出版文化や学術論文の価値とは何か
  - ○ ほとんど引用されない数多くの学術論文の価値（→ impact factor の価値）
  - ○ めったに売れないあまたある専門書の価値
- 全体と個
  - ○ 少数のヒットが全体を支えているとしても，どの少数かは，
    - 時間を経ないと分からない（社会が行き詰まっている場合）
    - 環境が変われば変わる（遺伝子の多様性の意義）．

☆「ヒットが無くなったから本屋を潰し大学を潰す」のではなく，できるだけ多くを支えるべき．

ランキングによるヒット集中の弊害を切り崩す

賞を作って隠れた名作を掘り起こそうとすると，受賞作だけがヒットになって終わる皮肉を打破する必要

- 売れないものにも，たまに，**ランダムに光を当てる**ことが重要ではないか？
- Stochastic ranking process
  - ○ 論文を机に積み上げて誰かが引用したら引っ張り出して眺めて一番上にぽんと置く．上から何部目かという順位．
  - ○ めったに売れない本の順位が乱高下する究極の順位定義．真価は長い時間かけねば分からないから乱高下して当然．**単純・自然かつ客観的な順位！**
- **出版関係者向け提案**：年１度大賞を発表する代わりに，たとえば本屋大賞の得票分ずつ玉を入れた福引きのガラガラを365回まわして順番を決めて，**「今日の１位」**を毎日発表．

文献

K. Hattori, T. Hattori, *Existence of an infinite particle limit of stochastic ranking process,* Stochastic Processes and their Applications (2008), to appear.
K. Hattori, T. Hattori, *Equation of motion for incompressible mixed fluid driven by evaporation and its application to online rankings,* preprint (2007).
K. Hattori, T. Hattori, *Mathematical analysis of long tail economy using stochastic ranking processes,* preprint (2008).

http://www.math.tohoku.ac.jp/~hattori/amazonj.htm

Google**検索キーワード　　服部哲弥**