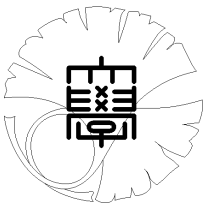
数理科学実践研究レター　2020–17 November 09, 2020

アクセスログを用いた
web ユーザーの行動予測定式化に向けて

by

小林 望



UNIVERSITY OF TOKYO

GRADUATE SCHOOL OF MATHEMATICAL SCIENCES

KOMABA, TOKYO, JAPAN

アクセスログを用いたwebユーザーの行動予測定式化に向けて

小林　望[1]（東京大学大学院理学系研究科物理学専攻）
Nozomu Kobayashi (Department of Physics, The University of Tokyo)

概 要

本研究ではマクロミル社が保有するwebユーザーの膨大なアクセスログからどのように各々のユーザーの行動を予測するかを考察した. 具体的には行動予測のモデル化を２つの方法で与えた.

1　はじめに

インターネットは我々の生活に必要不可欠なものとしてその重要性は日々増している. 各個人がインターネット上でどのように活動しているか, という情報はビッグデータとしてインターネットサーバー等に記録されている. 特にマクロミル社は日本全国にいるモニターらがどのwebページにいつアクセスしたかというログを保有している. このアクセスログから各モニターが次にどのようにweb上で行動するかを予測することはマーケティング上非常に重要である.
本レターではweb空間をグラフ理論を用いて定め, マクロミル社の保有するアクセスログデータを用いたユーザーの行動予測に応用するための定式化を考察する.

2　ランダムウォークを用いた定式化

本章からは以上で述べたユーザーの行動予測を具体的なモデルとして定式化することを目指す. 非常に一般的に考えると, 我々が目指すべき最終的な目標は全ての時刻tにおいてあるユーザーのwebページ遷移を予測することである. すぐに明らかなようにこの問題は非常に難しい. そのため, 全ての時刻を問題にするのではなくある時刻に着目する. 次に仮定としてユーザーのwebページ遷移は各時刻で完全にランダムに起きるものとする. すると与えられたweb空間における典型的な時間スケールはほぼなくなってしまい, 我々が注目すべき時刻とは唯一$t \to \infty$の場合, つまり長時間極限を取った時である. そこで以下では, あるweb構造が与えられたときにユーザーが長時間極限でどのwebページにアクセスしているかという問題をグラフ理論の言葉を用いて定式化を行う. このような定式化は**ページランク**と呼ばれ, 後で述べるようにGoogle等で使われているものである. 本章ではグラフ理論から始めてページランクのレビューとその実際のデータへの応用について述べる. ページランクについての詳細は [1] を参照のこと. 以下断りの無い限りグラフは有限グラフであり, 時間は離散時間を指す.

2.1　グラフ理論とweb空間

まず,web空間をグラフ理論の言葉で定義したい. そこで以下のようにこの空間を特徴づけする:

- 空間の一点一点は各々のwebページを表す.
- あるwebページaから別のwebページbへリンクが貼られていた場合, この空間では点から点への矢印として表す.
- 各ユーザーがwebページ遷移を行うことは, この点上を矢印にそって移動することに対応する.

これらの特徴づけからweb空間を**有向グラフ**として定義することができる.

[1] nozomu.kobayashi@ipmu.jp

定義 1 頂点の集合を V, 辺の集合を E, とするとき $G=(V,E)$ をグラフと呼ぶ. また各辺 e に対し向きがあるとき有向グラフと呼ぶ.*web* 空間の場合,V = { 与えられた *web* ページ v の集合 }, E = { 各 *web* ページ間のリンク e の集合 } である. この時,

$$V_v = \{\text{ ページ } v \text{ からリンクが飛んでいるページの集合 }\} \tag{1}$$

$$B_v = \{\text{ ページ } v \text{ へリンクしているページの集合 }\} \tag{2}$$

$$|v| = \{\text{ ページ } v \text{ から出ていくリンクの数 }\} \tag{3}$$

と定義する.

我々が知りたいのはユーザーがある時刻 t でどの web ページにいるかだが, これはユーザーが各 web ページに滞在している確率として表される.

定義 2 与えられた *web* 空間においてある時刻 t で *web* ページ v_i にユーザーが滞在している確率を $p_i(t)$ と表す. 次の時刻 $t+1$ においてユーザーは v_i に貼られたリンクにそって別の *web* ページに移動する. どのページに移動するかは等確率であるとする. すると,$p_i(t)$ は次の漸化式を満たす.

$$p_i(t+1) = \sum_{v_j \in B_{v_i}} \frac{1}{|v_j|} p_j(t) \tag{4}$$

さらにユーザーが *web* ページに長時間経過後滞在している確率は極限値として与えられる.

$$p_i = \lim_{t\to\infty} p_i(t) \tag{5}$$

実際に, このように定義した極限値は**ページランク**と呼ばれ, その web ページの重要度を測る指標になっており, 検索エンジン Google などでも使われているものである. 次にどのようにして極限値, ページランクを計算するかを述べる.

2.2 グラフの定量化と極限値

上で定義した極限値を計算するには,web 空間を表すグラフそのものを定量的に扱うことが便利である. そのためにグラフそのものを行列を用いて定義する.

定義 3 与えられたグラフ $G=(V,E)$ に対応する隣接行列 A を以下のように定める.

- 各々の頂点 v に順番をつける.*(1,2,* $\cdots$ *N)*
- $N\times N$ 行列 $A=(a_{ij})$ に対し,

$$a_{ij} = \begin{cases} 1 & (\text{頂点 } i \text{ から } j \text{ にリンクがある. }) \\ 0 & (otherwise) \end{cases} \tag{6}$$

と定める.

このように定めた隣接行列 A はグラフの情報を全て含んでいる. この隣接行列からページランクを計算するには, 新たに確率遷移行列 Q を,A の行ごとの和が 1 になるように規格化したものを考えればよい. すると, ユーザーの滞在確率ベクトル $\mathbf{p}(t)=(p_1(t),p_2(t),\cdots,p_N(t))^T$ の満たす漸化式は

$$\mathbf{p}(t+1) = Q\,\mathbf{p}(t) \tag{7}$$

となり, 時刻 n における確率分布は $\mathbf{p}(n)=Q^n\,\mathbf{p}(0)$ と計算することができる.
さて, ではどのようなグラフを与えた時でも確率分布は極限値を持つのだろうか. 実際, 以上のように定まる確率分布が極限値を持つためには次の 2 つの必要条件を満たさなければならない.

1) グラフが強連結　$\Leftrightarrow$ 隣接行列が基底によらず既約

2) グラフが非周期　⇔ 既約な確率遷移行列の対角成分が少なくとも一つ正

一般に, 適当なグラフを考えるとこれらの条件を満たさないことがわかるが, 適当に確率遷移行列を変形してあげることで条件を満たすようにすることができる.Google のページランクでは実際にそのような操作を行っている. 具体的にはテレポーテーション行列と呼ばれる補助行列を考える.

$$T = \frac{1}{N}\begin{pmatrix} 1 & \cdots & 1 \\ \cdots & \cdots & \cdots \\ 1 & \cdots & 1 \end{pmatrix} \tag{8}$$

T と先に与えた Q の凸結合を考えると, これは極限値を持つことが知られている.

2.3　マクロミルデータへの応用

これまで web 空間をグラフ理論を用いて定量的に定式化し, 極限値の計算手法を与えた. ここでは最後にマクロミル社の保持する実際のアクセスログデータをどのように応用できると期待されるか, その一例を述べる. アクセスログデータは, 性別, 年齢や年収等の様々な属性のついた個人ごとのデータの集合になっている. したがって属性ごとに区切った各ユーザーごとにグラフを形成することができる. このとき, 同じページに何回もアクセスしている場合, グラフの辺を等確率とするのではなく重み付けに変更することでこの差異をグラフに反映させることも可能である. そのように属性ごとに区切ったグラフごとに先程与えた極限値からページランクを求める. 例えば属性ごとに別々に与えられるグラフのページランク同士を比べることでページの重要度が属性によってどのように異なるかを見出せば, マーケティング上有用な情報をページランクから取り出すことが可能であろう.

3　拡張した web 空間による定式化

先の章で考えた web 空間上のユーザーの行動予測モデルは,web ページのみをグラフとして考え, ユーザーはそのグラフ上を動くものと考えていた. 本章ではこの考えを拡張させ, ユーザーそのものもグラフ上の頂点とするモデルを考察する.
具体的には, 以下のように時間変動するグラフを考える.

定義 4 拡張された *web* 空間を以下のグラフ $G(t) = (V, E(t))$ として定める*:*

- $V = (V_{web}, V_{user})$. ここで V_{web} =$\{$*web* サイトの集合 $\}$,V_{user} =$\{$*web* ユーザーの集合 $\}$.
- V の要素それぞれに $i = 1, \cdots N$ まで番号を割り振る.
- $e_{ij}(t) \in E(t)$ は時刻 t で頂点 i と j の間に以下の関係があるときに存在する.

  web サイト同士の場合　リンクが貼られている

  web サイトとユーザーの場合 ユーザーがサイトに滞在している

  ユーザー同士の場合 連絡先を知っている

以上で定義した web 空間においては, ユーザーが時刻 t であるサイトに滞在しているという状態はグラフ上の辺として表現される. 付随して, ユーザー同士の繋がりを辺として表現することができる.
このように定義したグラフに対応する隣接行列は時間に依存する行列として表される:

$$A(t) = \begin{pmatrix} B & C(t) \\ C(t) & D(t) \end{pmatrix} \tag{9}$$

ここで B は web サイト間の隣接行列,$C(t)$ は web サイト, ユーザー間の隣接行列,$D(t)$ はユーザー間の隣接行列である. また, 簡便のために web サイト間のリンクは時間変動せず一定であるとした. 次に隣接行列 $A(t)$ を以下で定める行列たち A_k を用いて以下のように分解する:

$$A(t) = \sum_k a_k(t) A_k \tag{10}$$

ただし集合 $\{A_k\}$ は**定義 4** を満たすグラフの隣接行列の集合である. いまグラフの時間変動依存性をこの行列 A_k で展開した際の係数に押し込めることができた. この時間変動する展開係数をそのグラフ A_k が時刻 t で実現する確率あるいは実現可能性を表す指標と思いたい. 上の分解では $a_k(t)$ について何の条件もついていなかったので, 以下のように正の 0 以上 1 以下の値に規格化する.

$$p_k(t) = |a_k(t)|^2 / Z(t), \quad Z(t) = \sum_i |a_i(t)|^2 \tag{11}$$

このように定義した $p_k(t)$ はグラフ A_k が実現する確率とみなすことができる.
以上のようにある時刻 t における web 空間を表す行列が与えられれば, そこから各実現可能なグラフの確率を求めることが原理的に可能である. したがって実際にどのようにしてグラフの時間発展を記述するかが問題になる. ここでは, グラフのダイナミクスをモデル化するためにグラフの頂点同士に”相互作用”を導入する. その作用は隣接行列 $A(t)$ に作用する $N \times N$ 行列 H として表される. 例えば, 頂点 i と j の間に大きさ h_{ij} の相互作用が働くとすると,$H = (h_{ij})$ である. ここで”相互作用”を記述する H は手で決めなければならないが, 今の例では例えば

- サイト, ユーザー間ではコンテンツの魅力度
- ユーザー間では SNS の友達リストやフォロー・フォロワー関係など

などを考慮して決めればよい.
相互作用行列 H のみによってグラフの時間発展が決まると仮定したとき, グラフの時間発展はどのような方程式に従うだろうか. 今まではグラフ理論的に考察を進めてきたが, グラフの頂点を格子点だと思い, それぞれの点に力学的自由度が内在しているとみなすと相互作用行列 H によって定まる時間発展は物理学における量子多体系のそれと類比して考えられないだろうか. つまり, 各グラフの頂点に電子やスピン自由度があり, 相互作用行列 H はこの系のハミルトニアンとみなすということである. ここではこのアナロジーを積極的に利用することで, グラフの隣接行列 $A(t)$ が従う方程式がハイゼンベルグ表示のシュレーディンガー方程式であるとして定式化する. つまり,$A(t)$ は

$$i\frac{d}{dt}A(t) = [A(t), H] \tag{12}$$

なる微分方程式に従うとする. この方程式の下では行列 $A(t)$ の時間発展は

$$A(t) = e^{iHt} A(0) e^{-iHt} \tag{13}$$

として与えられる. 以上のように, 相互作用行列を手で与える必要があるがこれを決めればグラフの時間発展を決定することが原理的に可能であり, そこからありうるグラフの実現可能性を評価することが出来る.

4 終わりに

本レターでは web ユーザーと web サイトからグラフ理論を用いて web 空間を定め, ユーザーの行動予測をどのように定式化するかを 2 つの方法で考察した. 1 つ目の手法では, ユーザーはグラフ上を動く点として表されユーザーの行動を逐一予測する代わりにページランクと呼ばれる指標を利用した. 2 つ目の手法では, ユーザー自身もグラフ上の頂点として表し, グラフ自身が時間変動するものとして定義した. ここで物理学からインスパイアされた概念である相互作用行列 H を導入することでグラフの時間発展をモデル化した.
今後の問題として具体的に以下のようなものが考えられる:

- 一つ目の手法について:
  - ページランクが極限値と定義されない場合 (極限が振動してしまうような場合) にもいくらか定量的な指標として扱えないか？
  - 実際のマクロミル社が保有するデータは時間的に非常に疎密である. このアクセスログから具体的にどのようにグラフを構成するか？
- ２つ目の手法について:
  - toy model として小さな web 空間を考え実際に H を構成して, その時間発展を調べる. 実際のマクロミルのアクセスログを再現するかどうか検証する.

5 謝辞

課題を提供していただいたマクロミルの皆様，研究にあたり多くの有益な助言をいただいた田中雄一郎氏に心より感謝を申し上げます.

参考文献

[1] A.N.Langville, C.D.Meyer, 『Googe PageRank の数理』岩野 和生, 黒川 和明, 黒川 洋訳 (共立出版, 2009)