

マルコフモデルを使用した Web ランキング

3X-06

赤津秀之

山名早人

早稲田大学理工学部情報学科

1 はじめに

Web を対象としたサーチエンジンでは、検索結果の並び順が重要な要素となってくる。

検索結果はランキングアルゴリズムによって検索語に適したものから順に並び替えられる。代表的なものに Google [3] で用いられている Pagerank[1]などがある。本研究ではこれに対し、マルコフモデルを用いた手法[2]について考える。[2]においては概念のみが記されていたが、本研究では、その中のパラメータの意味について調べた。

2 マルコフモデルで web 巡回を抽象化

仮想ユーザによる Web 巡回を一つのページから別のページへの移動と考え、これをマルコフ連鎖としてモデリングしていく。

検索質問 q において $R = \{r_1, r_2, \dots, r_n\}$ を検索エンジンによって見つけた Web ページの集合とする。そして、Web 巡回をしている仮想ユーザに関して、時刻 t において r_i の web ページをブラウズしている確率を $p(t)$ とし、そこから r_j にジャンプする確率を p_{ij} とする。つまり、ユーザが Web 巡回する過程をマルコフ連鎖として抽象化することができる。

定理 1

均質マルコフ連鎖の振る舞いは、その初期の分配ベクトル $p(0)$ および推移確率行列 $P, p(t) = p(0)P^t$ によって決定することができる。

時刻 t において仮想ユーザが r_i のサイトをブラウズしているとき、 $t+1$ において、仮想ユーザは次の 4 つのうちのどれかの行動をとると考える。

- 同じページをブラウズしつづける (α)
- ハイパーリンクをクリックして別のページにジャンプする (β)
- ブラウザの Back ボタンを押して前のページに戻る (γ)
- URL を直接入力、またはブックマークを使うなど全く別のページへジャンプする。 (ε)

それぞれの選択肢について、仮想ユーザの傾向を $\alpha, \beta, \gamma, \varepsilon$ とし、

$$0 < \alpha, \beta, \gamma, \varepsilon < 1, \quad \alpha + \beta + \gamma + \varepsilon = 1$$

と定義し、特定のユーザにおける 4 つの定数を決める。

定義

R における Web ページの傾向行列を

$$U = (u_{ij})_{n \times n}, u_{ij} = \{\alpha \text{ or } \beta \text{ or } \gamma \text{ or } \varepsilon\}$$

として、Web ページ R における推移確率行列は

$$P = (p_{ij})_{n \times n}, p_{ij} = \frac{u_{ij}}{\sum_{j=1}^n u_{ij}}$$

で表すことができる。

定理 2

状態空間 S 、推移確率行列 P 、その初期分配ベクトル $P(0)$ において、複素均質のマルコフ連鎖 $\{x_t, t \geq 0\}$ は $t \rightarrow \infty$ において最終的に一定値に収束すると考える。つまり

$$\lim_{t \rightarrow \infty} p(t) = p_*$$

この最終的な分配ベクトル

$$p_* = (\pi_1, \pi_2, \dots, \pi_n)$$

は方程式 $p_* P = p_*$ の唯一の解であり、

$\pi_i > 0, \sum_{i=1}^n \pi_i = 1$ を満たす。

本研究において、ユーザが Web 巡回において、Web ページをブラウズする可能性はその価値に比例すると考える。時間を無限に考える場合、各 Web ページをブラウズする最終的な確率 p_* は各 Web ページの価値に依存するという考えに基づき、 p_* の各成分を web ページのスコアとする。

3 検証

$\alpha, \beta, \gamma, \varepsilon$ のパラメータを変化させたときに p_* がどのように変化するかを計算した。リンク構造として図 1 を使用した。

図 1 の作成に当たっては、被リンクと発リンクの差が影響するのではないかと考え、②と⑤、①と③と⑧あたりを状況の似ているページとして考えた。

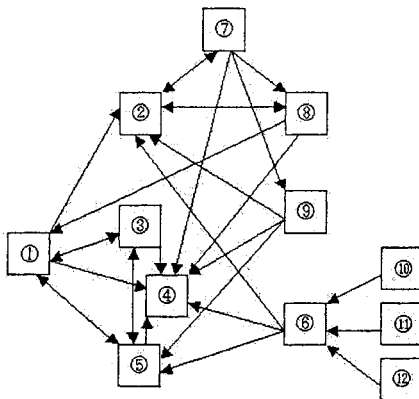


図1 リンク構造

今回、計算してみたパラメータの組み合わせは以下の4通りである。

- (1) $(\alpha, \beta, \gamma, \epsilon) = (0.6, 0.2, 0.19, 0.01)$
 (2) $(\alpha, \beta, \gamma, \epsilon) = (0, 0.55, 0.45, 0)$
 (3) $(\alpha, \beta, \gamma, \epsilon) = (0, 0.55, 0.44, 0.01)$
 (4) $(\alpha, \beta, \gamma, \epsilon) = (0, 0.7, 0.29, 0.01)$
 その結果次の通りになった。

	(1)	(2)	(3)	(4)
1位	④.111	④.137	④.139	④.176
2位	①.109	①.130	①.133	⑤.163
3位	②.109	②.128	②.129	①.131
4位	⑤.099	⑤.113	⑤.115	③.129
5位	⑥.097	⑥.101	⑥.100	②.128
6位	⑧.088	⑧.093	⑧.094	⑧.090
7位	⑦.087	⑦.090	⑦.091	⑦.063
8位	③.078	③.075	③.075	⑨.052
9位	⑨.077	⑨.073	⑨.074	⑥.051
10位	⑩.046	⑩.018	⑩.015	⑩.004
10位	⑪.046	⑪.018	⑪.015	⑪.004
10位	⑫.046	⑫.018	⑫.015	⑫.004

表1 各パラメータの組み合わせによるスコア

この結果を見ると、(1)~(3)では順位の変動は全くないが、(1)と比較して(2)、(3)では上位と下位の差が広がっていることから、 α の値を用いない方が明確なランキングを作るのに適しているのではないかと考えられる。次に(4)においては、リンクを辿って別のページにジャンプするパラメータ β を重視しているために、多くリンクが張られている④や⑤といったページのスコアが大幅に上がっているが、孤立している⑩~⑫にリンクが張られているだけの⑥は逆に順位を大きく落としている。

リンク状況の似たページの比較では、②と

⑤は(1)~(3)ではあまり差がなかったが(4)において順位が入れ替わった。⑤はもともとスコアの高い①や③からリンクが張られているのが大きいと考えられる。

①③⑧においては、③が大幅にスコアを伸ばしている。これもスコアの高い①や⑤からリンクが張られているのが大きいと考えられる。また、⑥の被リンクを増やすためのダミーサイトを想定した⑩~⑫は特に(4)においてスコアが急激に落ちている。やはり、こういったサイトのスコアを上げないためには γ よりも β を重視すべきではないかと考えられる。

4 おわりに

今回使用したパラメータの ϵ はランダムに飛ぶということを想定しているが、実際にはランダムに飛ぶことよりもブックマークなどで別のページへ移動することが多いので、そこを考慮したモデリングを考えていく必要があるのではないかと。

今回は簡単なリンク構造による結果であり、今後大規模な構造における実験が必要となる。この際計算量が膨大になるため、計算量に関して解決法を考えなければならない。

謝辞

本研究の一部は、文部科学省科学研究費補助金奨励研究(A)(課題番号: 13780255)及び早稲田大学特定課題研究(2001K-036)によるものである。

参考文献

- [1] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bringing Order to the Web', 1998,

http://www-db.stanford.edu/~backrub/page_ranks/ps

- [2] Dell Zhang, Yisheng Dong, 'An Efficient Algorithm to Rank Web Resources' Proc. of 9th International World Wide Web Conference, Amsterdam, Netherlands, May 2000.

- [3] Google
<http://www.google.com/>