主専攻実験 A 最終レポート

岡部 純弥

2022 年 7 月 25 日

概要

本課題では，Google の検索アルゴリズムとして非常に有名な PageRank アルゴリズム [1][2] の理論を理解し，これを用いた計算機実験を行った．実際に，日本国内の主要空港間の移動者数データに対して PageRank アルゴリズムを適用し，各空港の重要度を計算した．

1 はじめに

1.1 PageRank とは

PageRank とは，Brin, Page[1] によって提案された Google の検索システムで用いられているアルゴリズムである．PageRank では，Web ページ間のハイパーリンク関係を用いて，各ページの重要度を計算する．これは，**良い** *Web* **ページは別の良い** *Web* **ページからリンクされている**という考え方をもとに実現されている．Facts about Google and Competition [*1] によると，

> PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites

と確かに記載されている．またこの考え方は，論文の引用/被引用数ネットワークや共著ネットワークと非常に似ている．つまり良質な論文は，別な良質な論文からリンクされているという考え方である．実際に PageRank アルゴリズムを用いた論文の共著システムに関する研究として，Ma et al.[3], Ding et al.[4] などが挙げられる．

1.2 応用先

PageRank は，Web サイトの重要度付けの他にも，(ソーシャル) ネットワーク分析, 物理学, 化学, 生物学など多数の応用先がある．ソーシャルネットワーク分析の事例としては Bahmani et al.[5] などが挙げられる．また，PageRank の応用に関する総説論文としては Gleich[6]，Berkhin[7] が著名である．

2 PageRank

2.1 定義

ここでは基本的な[*2]PageRank のアルゴリズムを紹介する．

u をある Web ページとする．また，u **から**リンクする Web ページの集合を F_u，u **に**リンクする Web ページの集合を B_u とする．さらに，F_u の要素数 N_u [*3]，正規化するための定数 c を用いると，u のランク $R(u)$ は式 (1) によって定義される．

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{1}$$

$R(v)/N_v$ は，v のランクを F_u の要素数，すなわち v からリンクするページの総数で割ったものである．つまり，$R(u)$ は u にリンクするすべてのペー

[*1] https://web.archive.org/web/20111104131332/https://www.google.com/competition/howgooglesearchworks.html

[*2] Page et al.[2] の論文に基づいた

[*3] すなわち $N_u = |F_u|$

ジに対して $R(v)/N_v$ を計算し，その総和に c を掛けたものである．したがって，ランクの高いページからリンクされているページもまたランクが高くなる傾向にある．

式 (1) を別の観点から評価し直してみる．ある正方行列 A を考え，A の (u, v) 成分を

$$A_{u,v} = \begin{cases} 1/N_{u,v} & \text{if edge from u to v exists} \\ 0 & \text{otherewise} \end{cases} \tag{2}$$

と定義する．このとき，R をベクトルとして考えると

$$R = cAR \tag{3}$$

と表すことができる．これは R が A の固有ベクトルに他ならないことを示している．[*4]

しかし式 (1) の定義には少し問題がある．ある 2 つのページ u' と v' が相互にリンクしており，なおかつ他のどのページともリンクしない状況を考えてみる．さらに，別のあるページが u' あるいは v' にリンクしているものとする．このとき，ランクをうまく配分することができない．そこで，式 (1) の定義を，あるベクトル $E(u)$ [*5] を用いて式 (4) に再度定義し直す．

$$R'(u) = c\left(\sum_{v \in B_u} \frac{R'(v)}{N_v} + E(u)\right) \tag{4}$$

ただし，式 (4) において $||R'||_1 = 1$ [*6]を満たすものとする．式 (4) は $cE(u)$ の項によって正則化されているため，前述したような問題が起きることはない．以後，この定義を用いて議論を進める．

PageRank のより詳細な理論，およびその拡張に関しては，Page et al.[2]，Bianchini et al.[8]，Langville, Meyer[9] などを参照されたい．

[*4] さらにこのときの固有値は c である．

[*5] $E(u)$ は Web 上のランクのソースに対応している

[*6] $||R'||_1$ は，R' の L_1 正規化ノルムを表す．

3 計算機実験

本課題では e-stat [*7][*8] 上で入手できる，日本国内の主要空港[*9]間の令和 2 年 2 月の月間移動者数の旅客数を用いた．このデータでは，各 OD ペアに対する月間の旅客移動数が記載されている．[*10]

4 結果

空港	重要度
羽田	0.295
成田	0.083
新千歳	0.147
伊丹	0.101
関西	0.076
福岡	0.160
那覇	0.136

Table.1: 各空港の重要度

5 考察

空港	旅客数
羽田	20,606,398
成田	1,984,001
新千歳	6,436,335
伊丹	5,812,333
関西	2,051,220
福岡	6,485,437
那覇	6,588,217

Table.2: 令和 2 年度 年間旅客数（国内）

[*7] 政府統計の総合窓口

[*8] `https://www.e-stat.go.jp`

[*9] 東京国際（羽田），成田国際，新千歳，大阪国際（伊丹），関西国際，福岡，那覇の 7 空港

[*10] ただし，羽田-成田間，伊丹-関西間のデータは見つからなかったため，0 人として扱っている．

空港	重要度順位	旅客数順位
羽田	1	1
成田	6	7
新千歳	3	4
伊丹	5	5
関西	7	6
福岡	2	3
那覇	4	2

Table.3

6 まとめ

参考文献

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[3] Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.

[4] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

[5] Bahman Bahmani, Abdur Chowdhury, and Ashish Goel. Fast incremental and personalized pagerank. *arXiv preprint arXiv:1006.2880*, 2010.

[6] David F Gleich. Pagerank beyond the web. *siam REVIEW*, 57(3):321–363, 2015.

[7] Pavel Berkhin. A survey on pagerank computing. *Internet mathematics*, 2(1):73–120, 2005.

[8] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[9] Amy N Langville and Carl D Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.