# 利用履歴に基づくPageRankアルゴリズムの改良

京都大学大学院情報学研究科
向　亨○　成　凱　上林　弥彦

# 発表内容

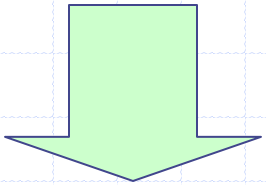
- 研究背景
- PageRankの紹介
- 改良案の提案
- 実装及び評価実験
- 実験結果の考察
- 将来研究及び結論

# 研究背景

＜研究目的＞

Web空間上におけるデータ量の増加

利用者の増加に伴うアクセス傾向の急速な変化



**履歴情報を利用した検索エンジンの提案**

- プロキシサーバから得られる
  アクセスログを解析

# 本研究の目的

## スコアリングアルゴリズムに着目

特定の基準によりWebページを定量的に評価
現在の検索エンジンにおいて重要な機能の1つ

＜従来のスコアリングアルゴリズム＞

- 文書解析によるアルゴリズム
  例) cosine distance、TF-IDF、ベクトル空間法
- リンク構造解析によるアルゴリズム
  例) HITS、PageRank

# PageRank

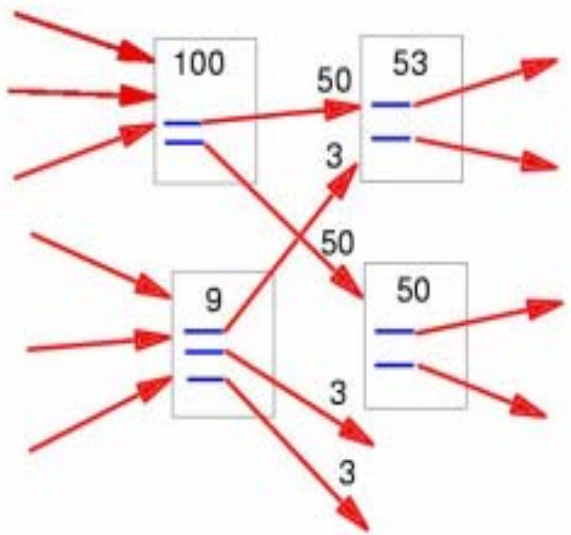
- L. Pageによって1998年に提案されたアルゴリズム
- 世界的に有名な検索エンジンGoogleで使用

＜基本概念＞

「重要なページは重要なページにリンクされる」

＜計算モデル＞

ページが持つ値を
リンク先へ等しく分配



# PageRankのアルゴリズム(1)

**行列式を用いて計算可能**

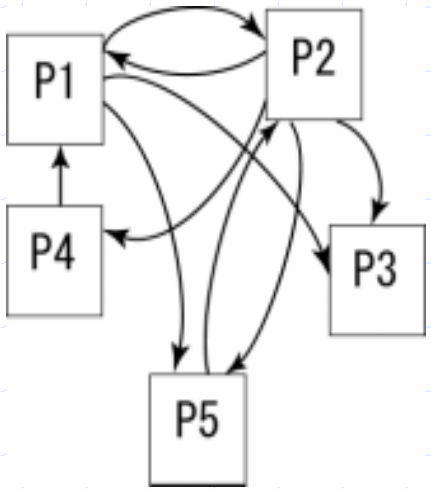
R = cAR ⇒繰り返し代入することで定常化

R:PageRankの値ベクトル

A:要素$A_{u,v}$
- $= 1/N_u$ (u,v間にリンクがある場合、$N_u$=ページuのリンク数)
- $= 0$ (無い場合)

行列Aの例)

$$\begin{pmatrix} 0 & 1/4 & 0 & 1 & 0 \\ 1/3 & 0 & 0 & 0 & 1 \\ 1/3 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 1/4 & 0 & 0 & 0 \end{pmatrix}$$



# PageRankのアルゴリズム (2)

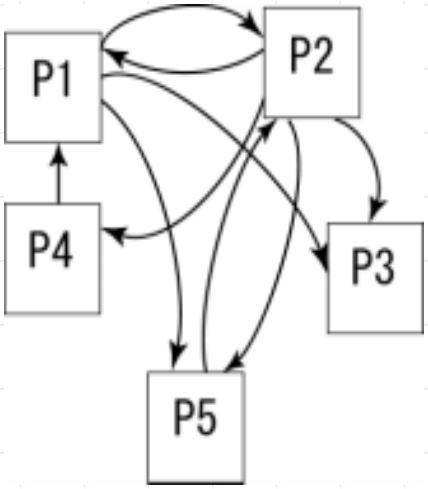
リンク以外のアクセス (ランダムジャンプ)
を考慮したモデル (値の発散を防ぐため)

$$R = c(A + E \times 1)R$$

E:全ての要素が1/(ページ数)のベクトル

*1*:全ての要素が1の正方行列

ベクトルE:
1/5
1/5
1/5
1/5
1/5



# PageRankの問題点

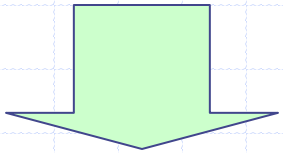
リンク構造という静的情報のみしか使えないために存在する問題点

1. リンクの等価性
2. リンク外アクセスの無視
3. 定常確率の無視

ログデータから得られる履歴情報を導入
WWW空間の動的な変化に対応できる
スコアリングアルゴリズムの提案

# 導入すべき履歴情報

1. Link Navigation History (LNH)
   **リンク個々のアクセス頻度**
   **リンクの持つ重要性を判断**
2. Page Jump History (PJH)
   **リンク以外によるページアクセス（ジャンプ）頻度**
   Bookmark**など重要な意味を持つアクセス**



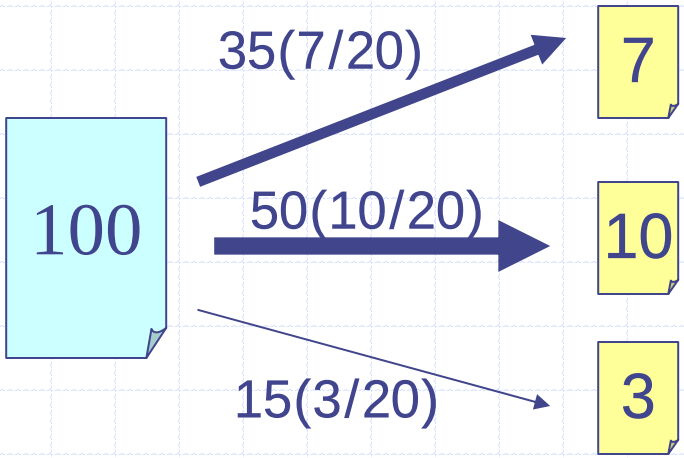
PageRank**に適用**

# 改良の概要

## 履歴情報を用いてPageRankを改良

### 1.LNHの導入

アクセス比に応じてリンク先に渡す値を決定




### 2.PJHの導入

回数に応じてランダムジャンプの確率を変化

### 3.定常確率の導入

一定比率（＝定常確率）だけ値を残す

# LNHの導入

R = cAR

A:要素$A_{u,v}$ = $hl_{u,v}$/ Σ $hl_{u,w}$(u,v間にリンクがある場合

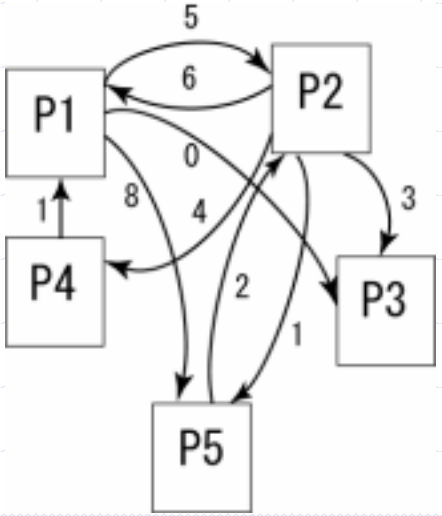
$hl_{u,v}$= u-v間リンクアクセス

w∈uからリンクしているページ)

= 0 (無い場合)

行列A:

| | | | | |
|---|---|---|---|---|
| 0 | 6/14 | 0 | 1 | 0 |
| 5/13 | 0 | 0 | 0 | 1 |
| 0/13 | 3/14 | 0 | 0 | 0 |
| 0 | 4/14 | 0 | 0 | 0 |
| 8/13 | 1/14 | 0 | 0 | 0 |

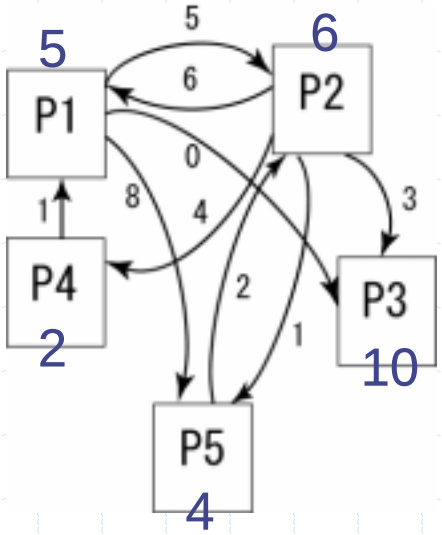


# PJH及び定常確率の導入

## ＜PJHの導入＞

$R = c(A + E \times 1)R$

ベクトルE: 5/27
6/27
10/27
2/27
4/27




## ＜定常確率の導入＞

$R = c(A + E \times 1 + C)R$

C:定常確率を表す正方行列
要素$C_{u,v}$はu=vの時のみ定数(滞在確率)

# プロトタイプの実装

ASTEM (京都高度技術研究所)
の協力で実際のプロバイダ
(Kyoto-Inet) のプロキシサーバ
で記録されたログデータを
テストデータとして利用
- 約130万 (html) ログ
- ログから得た約12万ページ
- 昨年7月のデータ


単語とランキングアルゴリズムを選択すると
上位10ページのリンクを出力

# 評価実験

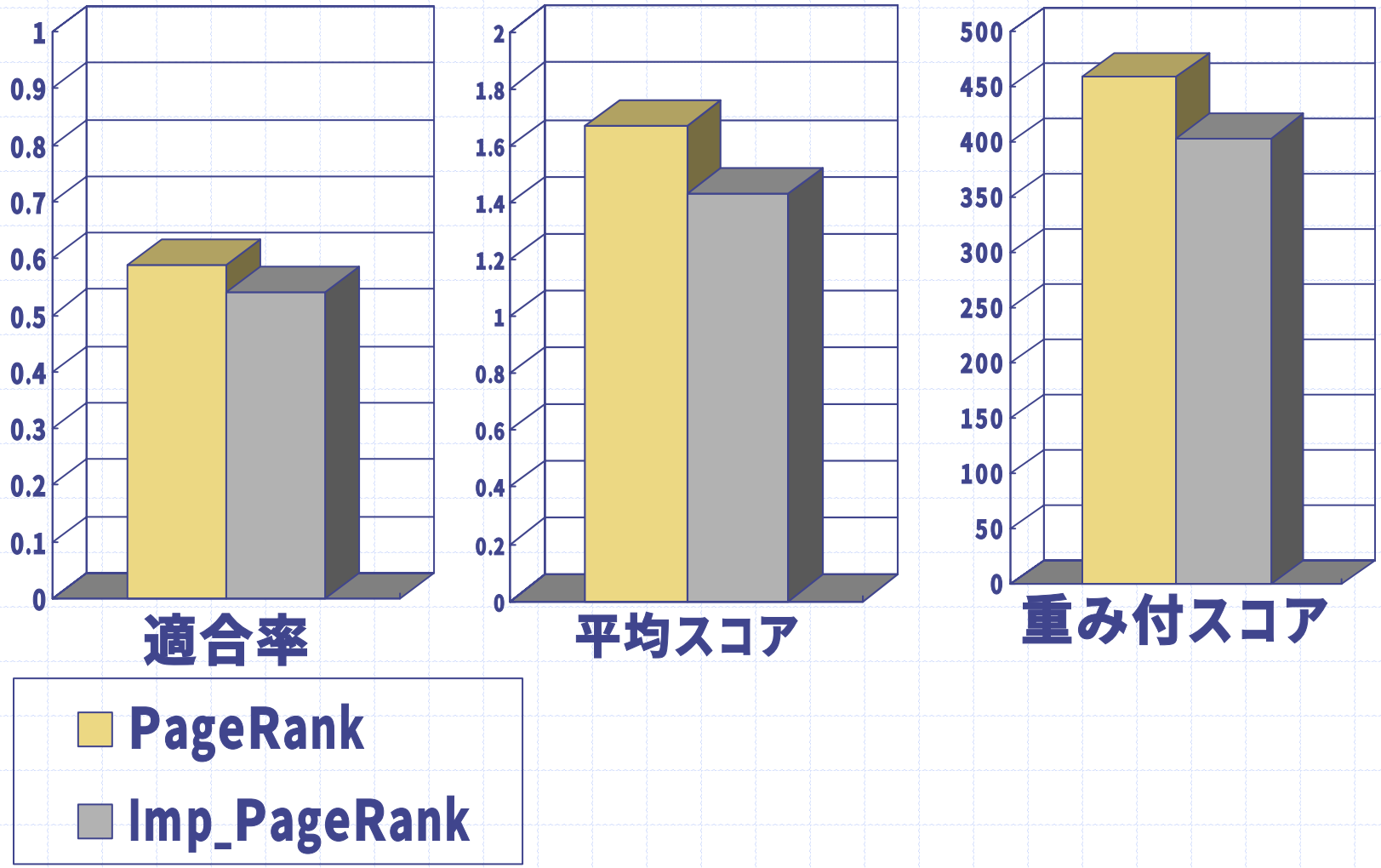
## ＜実験方法＞

- 被験者は実験側が提示した25個の単語から任意に5個の単語を選択し、2つのアルゴリズム全てに対して検索を1度ずつ行う
- 検索されたページを閲覧、0～5の6段階でそれぞれのページを評価

## ＜評価基準＞

1. 適合率（1以上の評価がついたページの割合）
2. 平均スコア（評価値の平均点）
3. 重み付スコア（上位ページに重みをかけたもの）

# 評価結果


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
適合率
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
平均スコア
500
450
400
350
300
250
200
150
100
50
0
重み付スコア
PageRank
Imp_PageRank

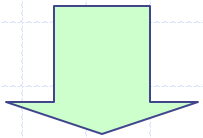
# 結果考察

## ＜実験結果＞

全ての基準においてPageRankが優位

## ＜原因考察＞

Imp_PageRankのみに出てきたページを検証



**時間に依存したページが頻出**

# 時間依存の事例（1）

## ＜ニュースページ＞

例）検索語「イチロー」


例）検索語「阪神」


内容がそのまま残っていれば高評価を受けるが、

**多くは削除されているか内容が変わっている**

# 時間依存の事例（2）

## ＜時期限定のイベント＞

例）検索語「ゆず」

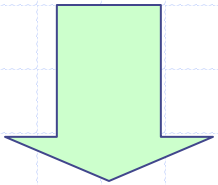

2002年2月

検索語


2001年7月

# 実験から得た結論

アクセス情報を利用した場合
データを**取得した時間に結果が依存**しやすい



以下の2点を徹底すれば有効なアルゴリズムとして活用できると期待される

1. より**新鮮な**ログデータ
2. **定期的な**データ収集

# 将来研究

## ＜履歴情報の利点＞

リンク、文書等の静的情報では扱えない
Web空間上の様々な動きに対応が可能

1. アクセス状況の**時間変化**
   特定トピックに関する流行のページを検索
   例）最近1ヶ月注目度が上昇している「ソルトレーク」
   に関するページ → オリンピックのページ
2. **特定期間**（日曜日、深夜等）のアクセス状況
   ある特定の時期における人気ページを検索
   例）3月の「倉敷」に関するページ → 音楽祭のページ

# まとめ

- 履歴情報を用いて
  PageRankアルゴリズムを改良
- プロトタイプを実装しアルゴリズムを評価
  - 結果は期待されたものと異なる
  - 原因の調査により時間依存が強さが判明
  - 新しいログデータによる改善が期待
- 将来研究

  履歴情報を用いた新たなモデルの提案
  履歴情報を用いた機能（時間変化、特定期間）追加