半構造データを用いた WWW リンク構造解析アルゴリズムの新提案

A New Algorithm for Analyzing the Link Structure of WWW Using Semi-structured Data

日本大学 文理学部 情報システム解析学科
谷 研究室　西村 有美子

目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

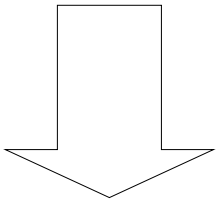
本研究の目的

WWW で情報を入手することは簡単
[検索エンジンの利用]
→ 目的のページが上位にくるとは限らない

本研究の目的

WWW で情報を入手することは簡単
[検索エンジンの利用]

→ 目的のページが上位にくるとは限らない



WWW リンク構造の解析を行い,トピックに
関係している web ページを正確に探し出したい

HITS アルゴリズムの改善手法を提案し
比較実験を行った

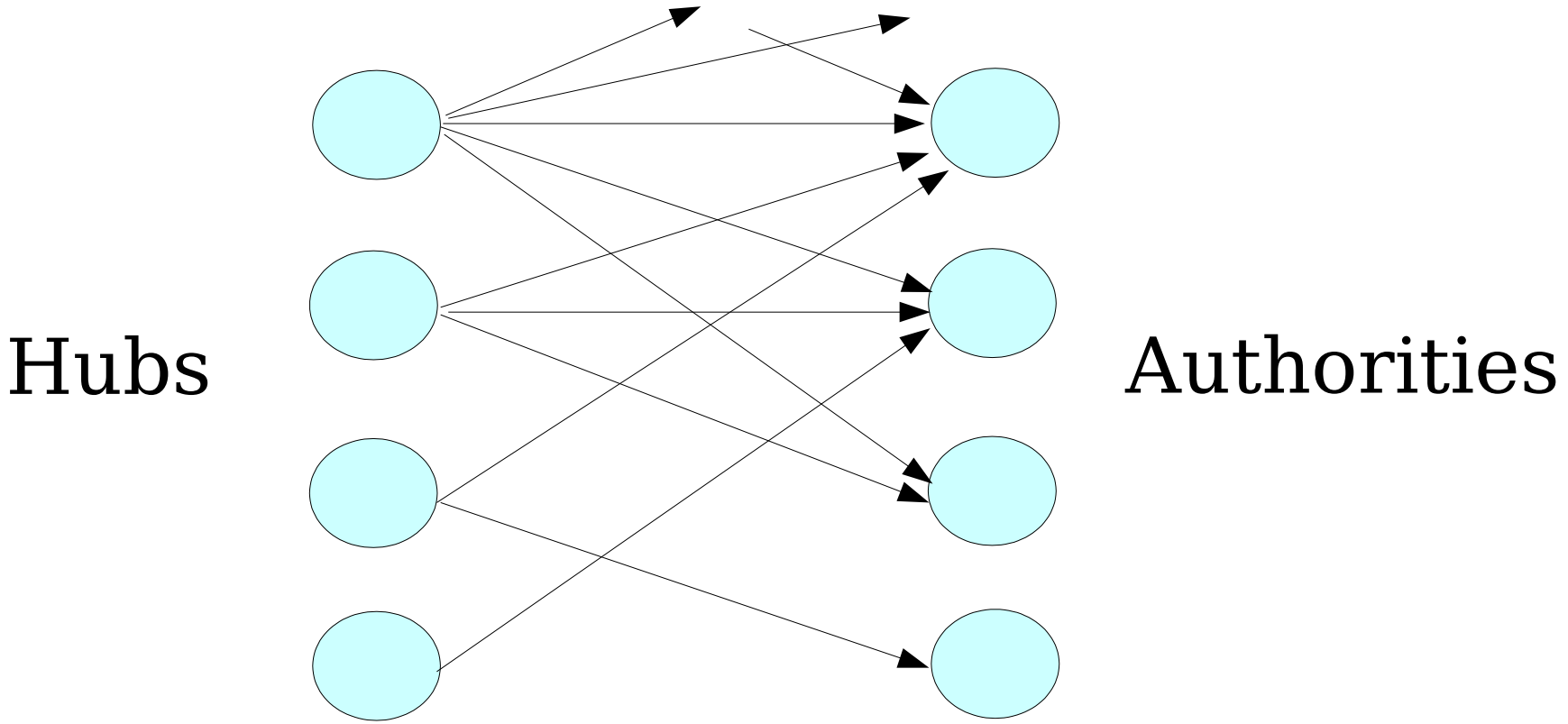
目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

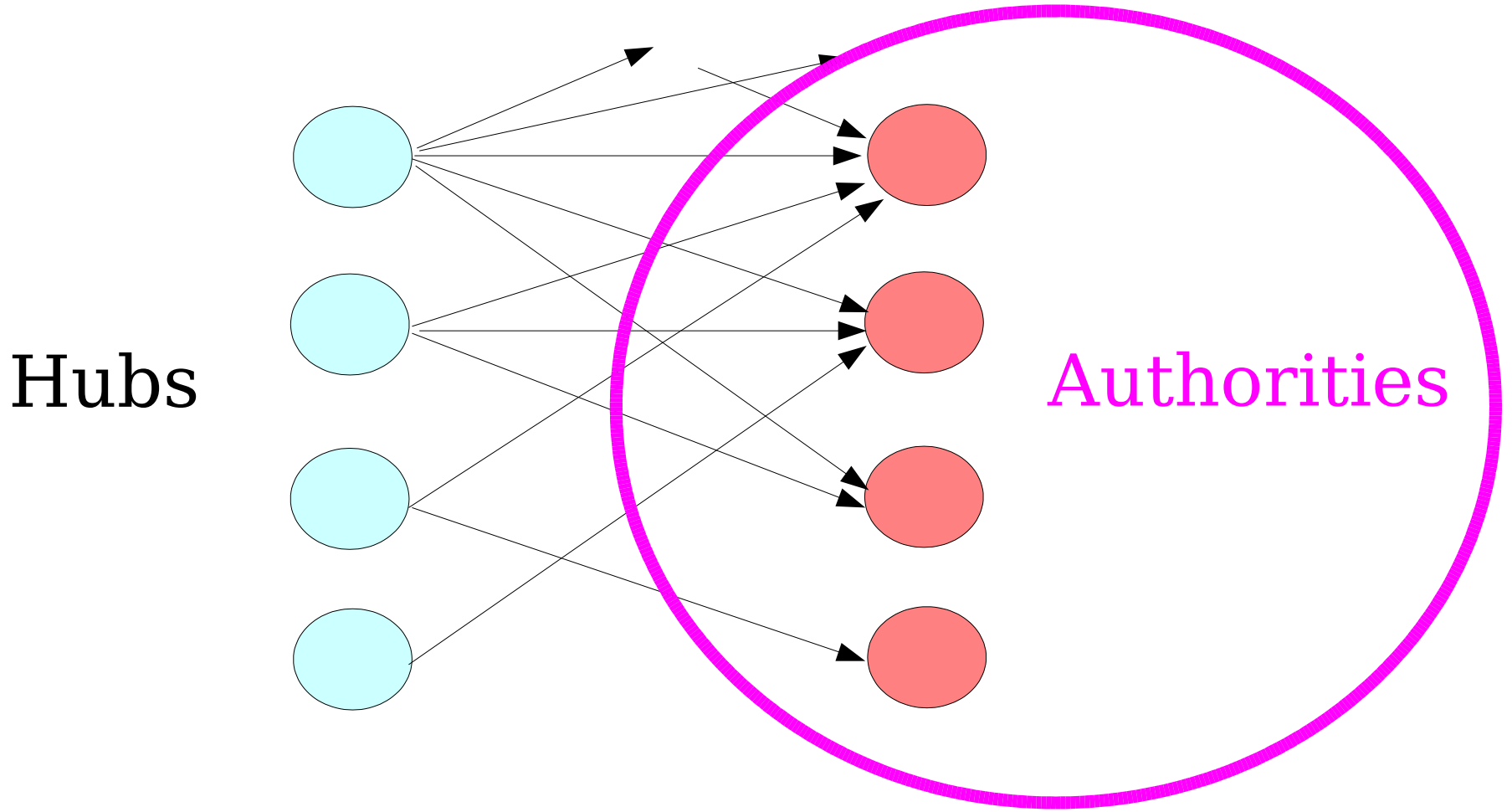
HITS アルゴリズムの特徴

* Jon M.Kleinberg 1997
* 各 Web コンテンツの内容には立ち入らず
  サイト間のリンク構造の解析のみで適切な情報
  を抽出する
* 適切な情報: Authority や Hub のページ集合

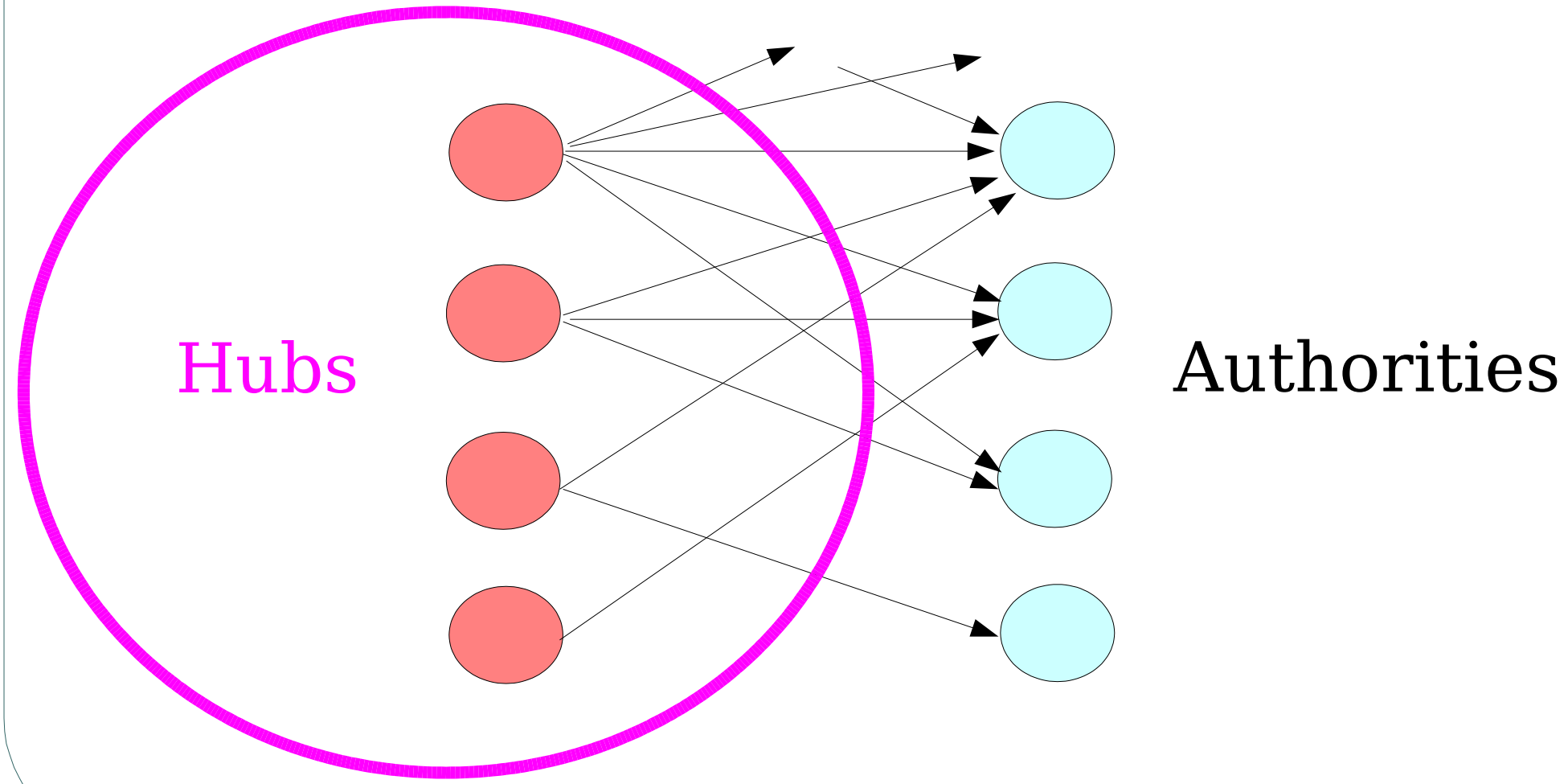
Authority と Hub



Authority と Hub



Authority と Hub



HITS アルゴリズムの手順

1 : root set の作成
2 : base set の作成
3 : Authority と Hub の重み付け

HITS アルゴリズムの手順

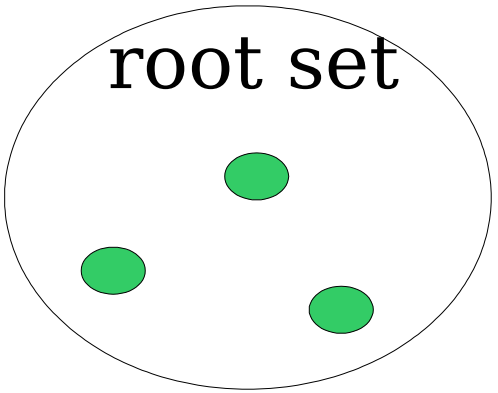
1：root set の作成

探したいトピックを検索エンジンにかけ
一定数 r 件の Web ページを収集し root set 作成

HITS アルゴリズムの手順

1：root set の作成

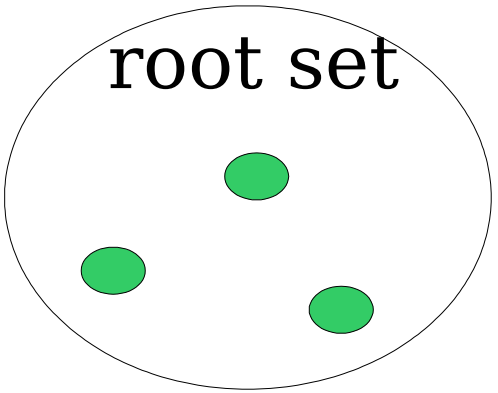
探したいトピックを検索エンジンにかけ
一定数 r 件の Web ページを収集し root set 作成



HITS アルゴリズムの手順

2：base set の作成

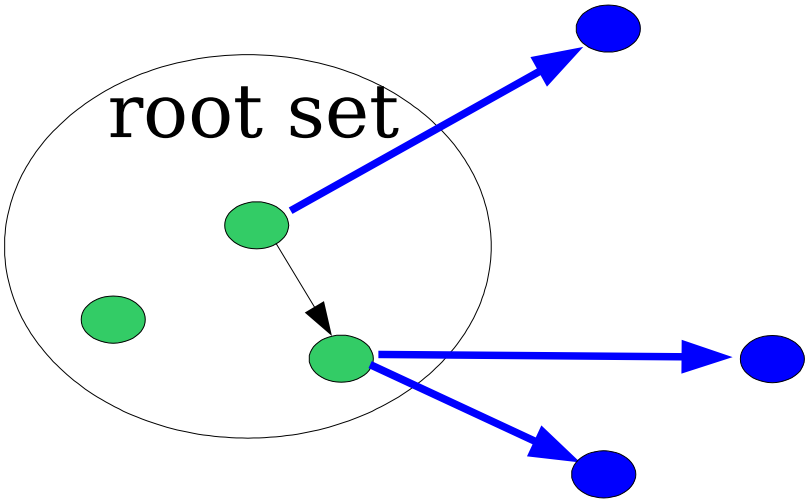
root set のページからリンクされている全てのページ
root set のページにリンクしているページ最大 d 件収集
root set に追加し大きさ n の base set 作成



HITS アルゴリズムの手順

2：base set の作成

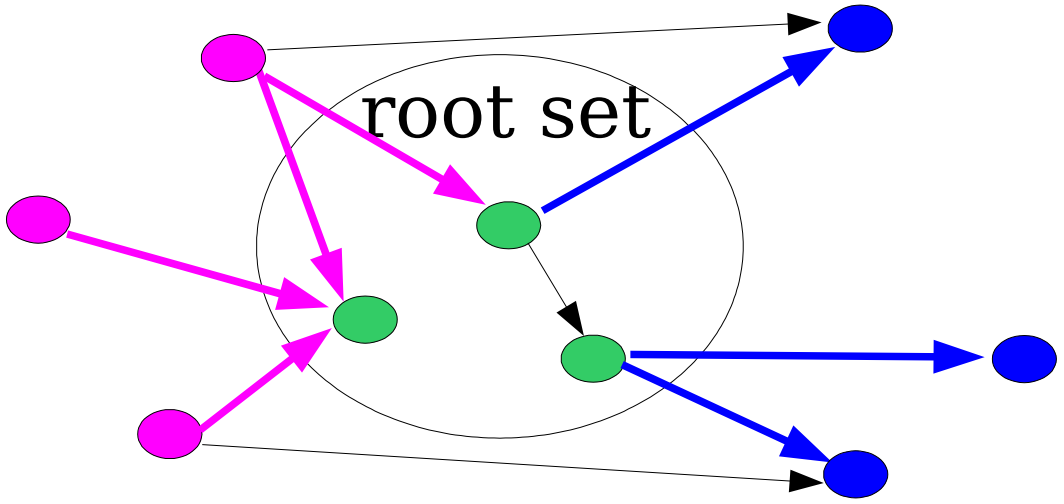
root set のページからリンクされている全てのページ
root set のページにリンクしているページ最大 d 件収集
root set に追加し大きさ n の base set 作成



HITS アルゴリズムの手順

2：base set の作成

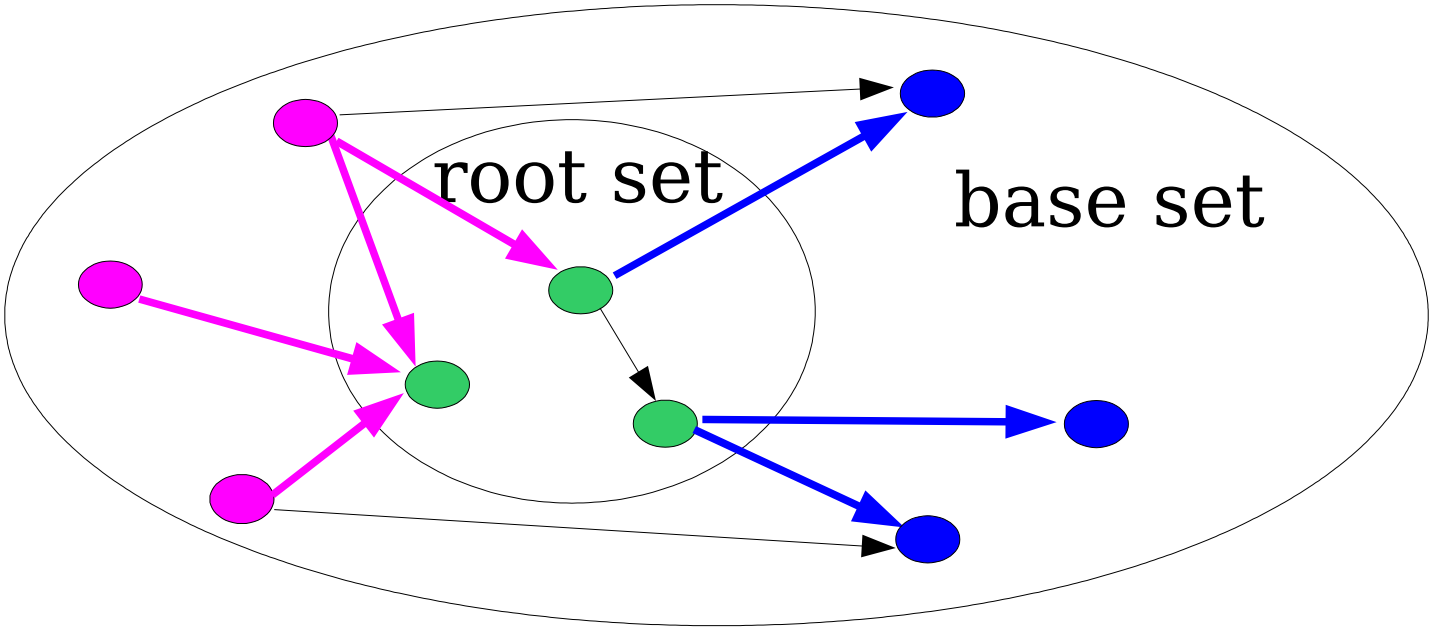
root set のページからリンクされている全てのページ
root set のページにリンクしているページ最大 d 件収集
root set に追加し大きさ n の base set 作成



HITS アルゴリズムの手順

2：base set の作成

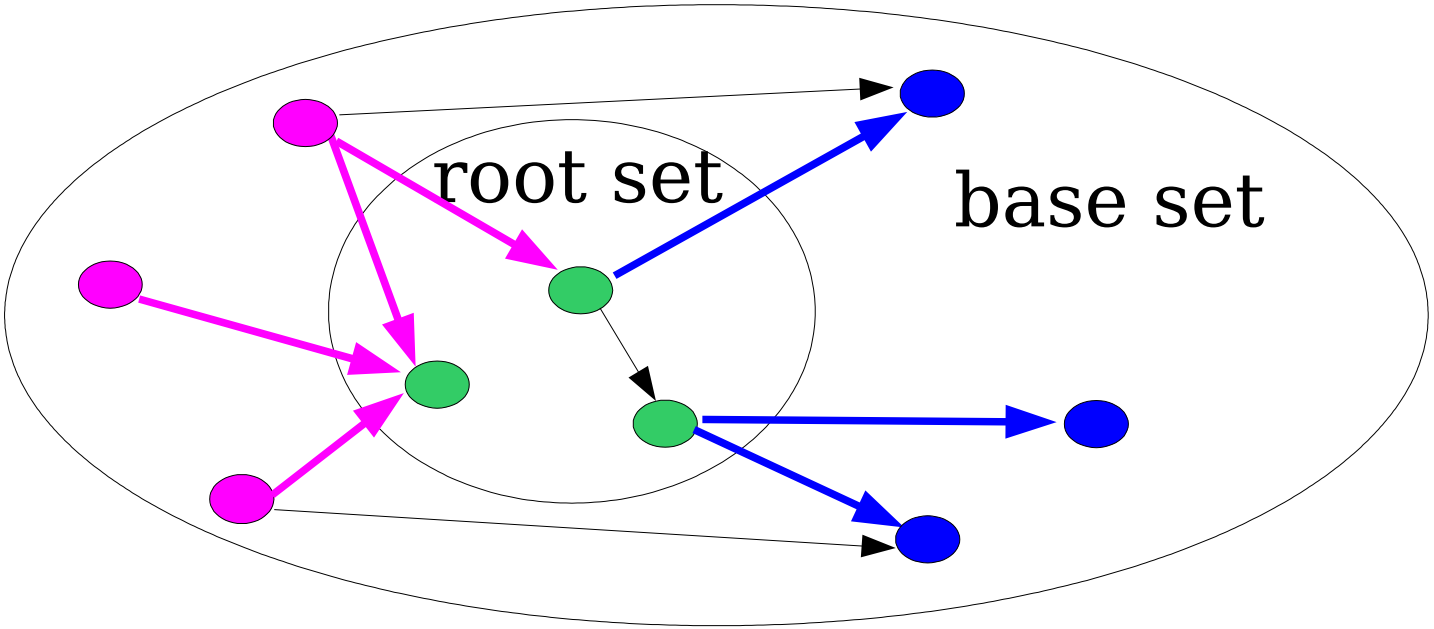
root set のページからリンクされている全てのページ
root set のページにリンクしているページ最大 d 件収集
root set に追加し大きさ n の base set 作成



HITS アルゴリズムの手順

3：Authority と Hub の重み付け

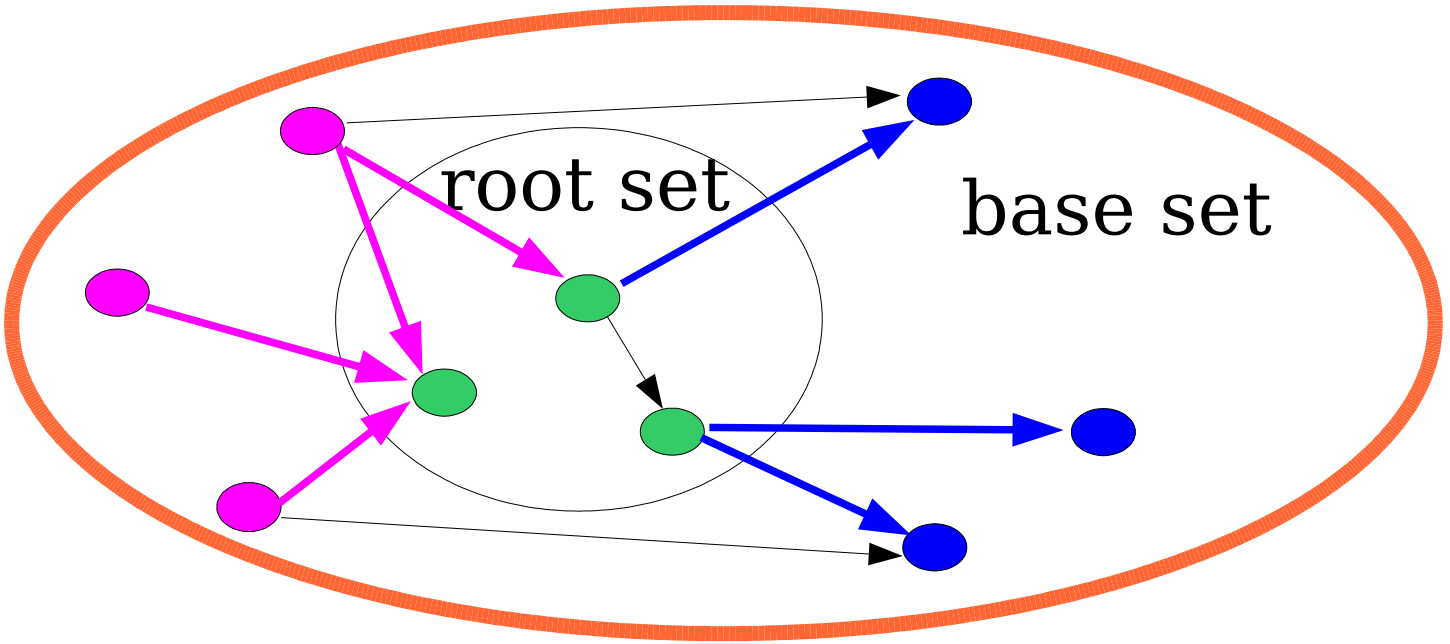
base set のページについて Authority と Hub に重みを付けていく



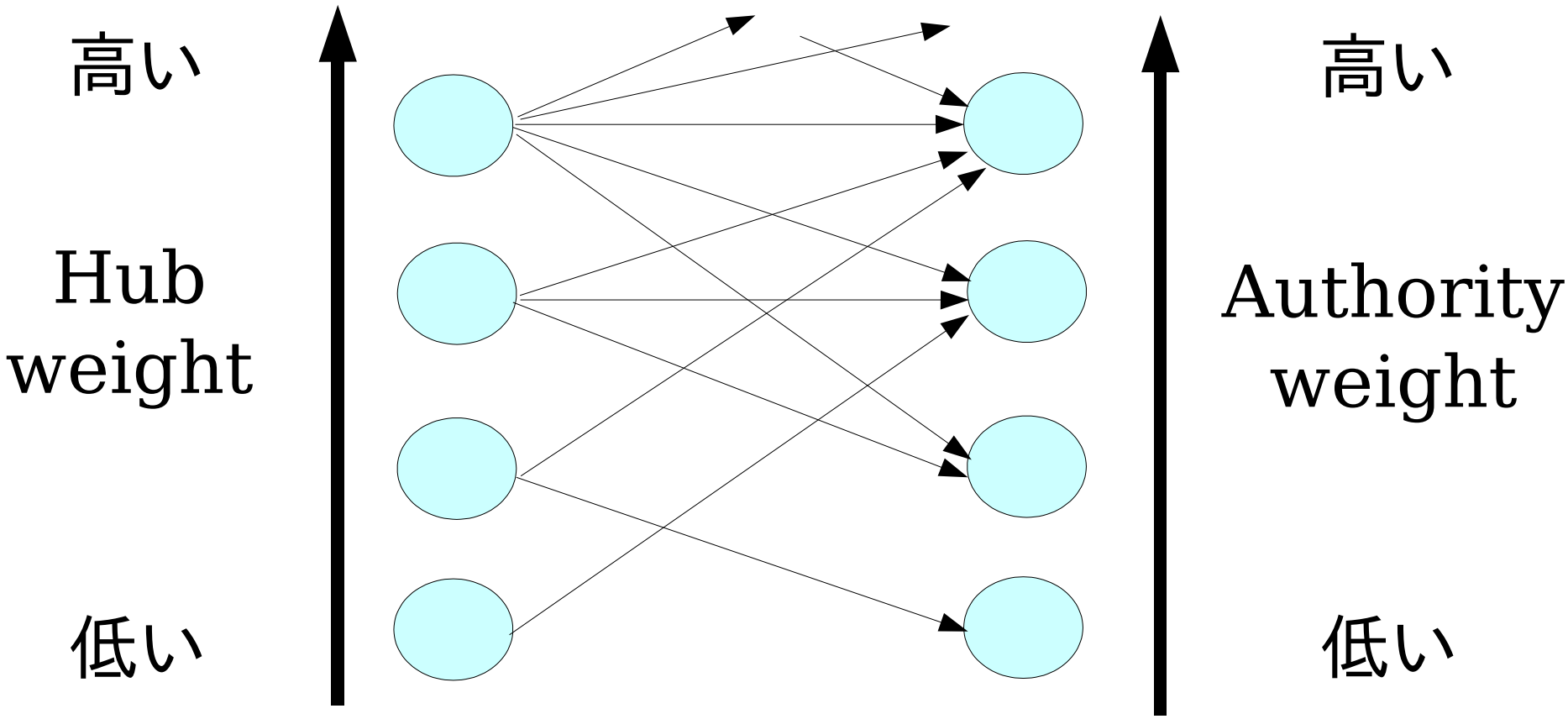
HITS アルゴリズムの手順

3：Authority と Hub の重み付け

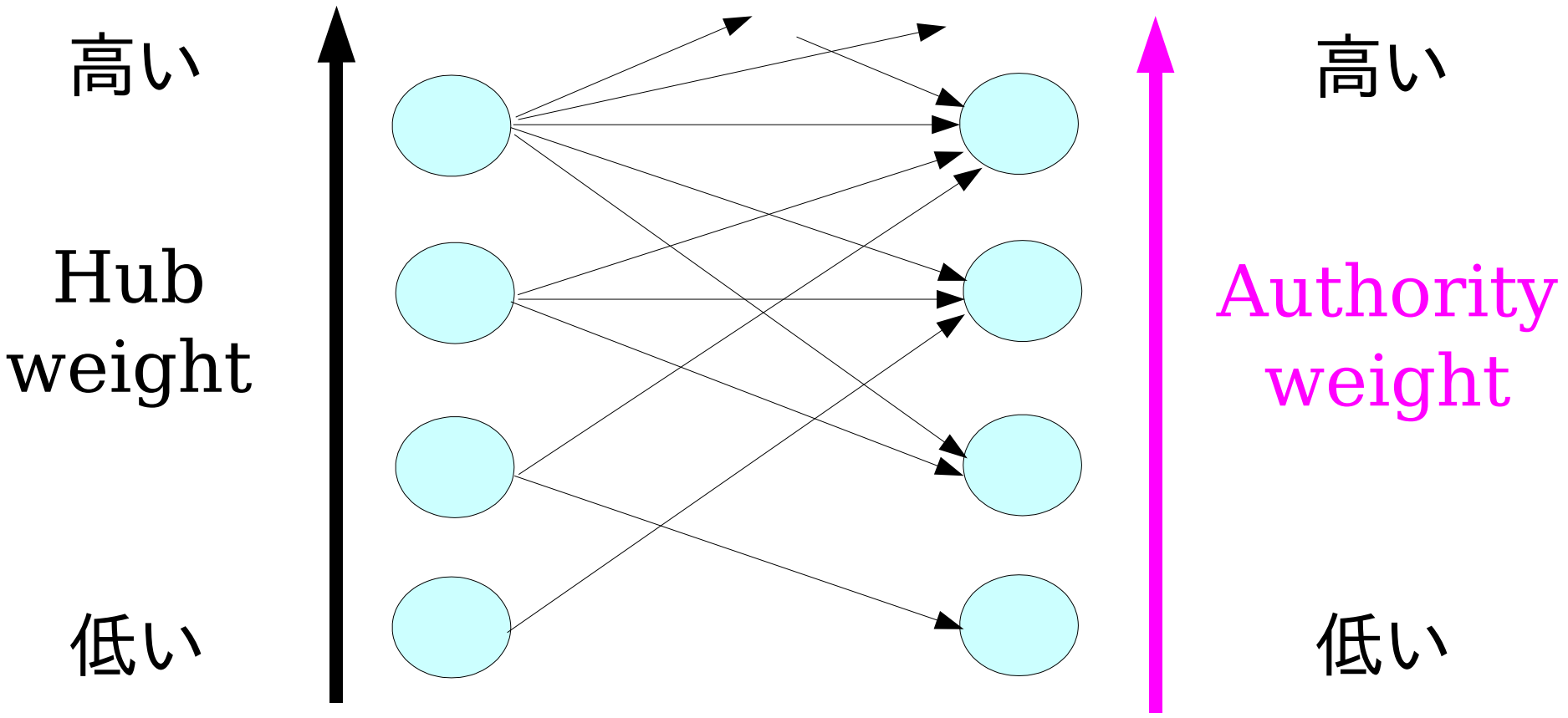
base set のページについて Authority と Hub に重みを付けていく



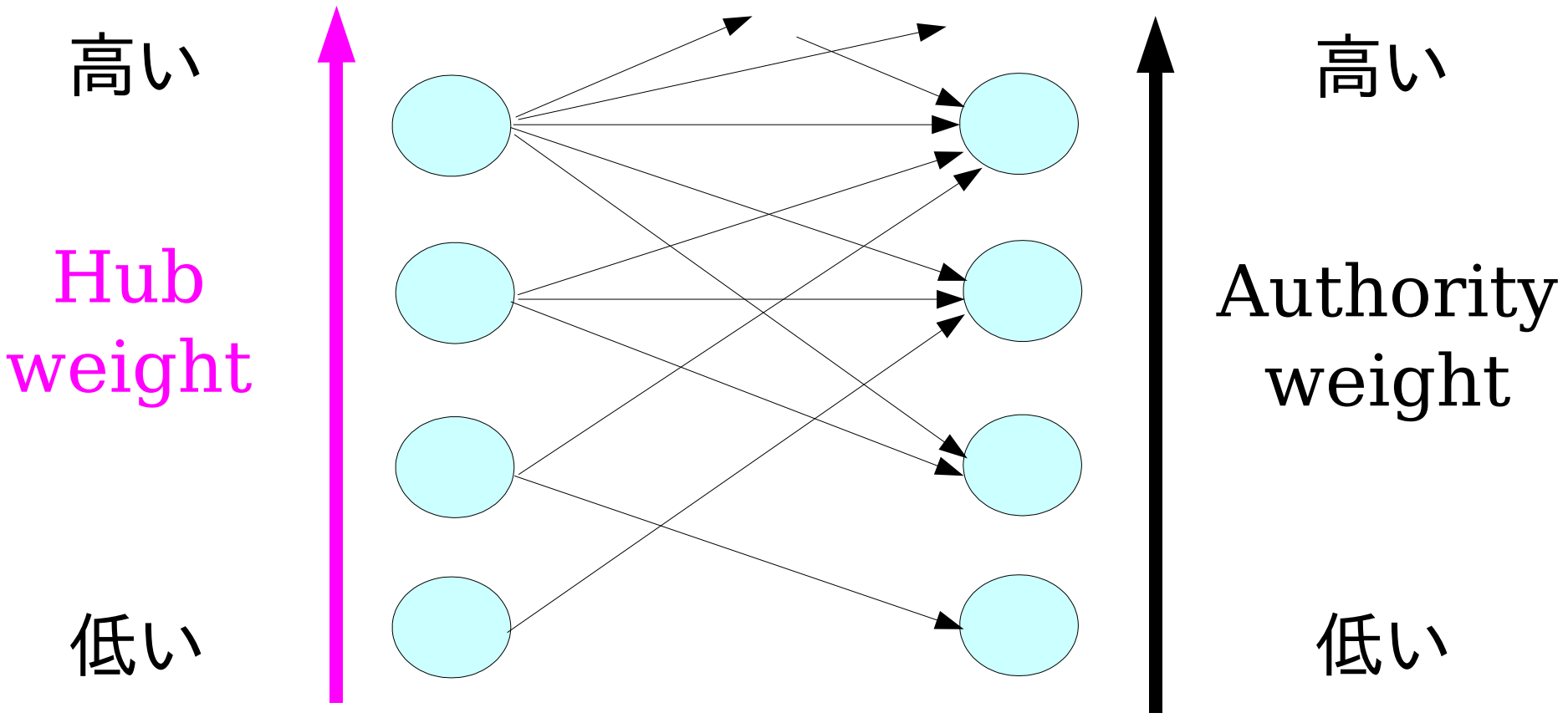
Authority と Hub の重み



Authority と Hub の重み

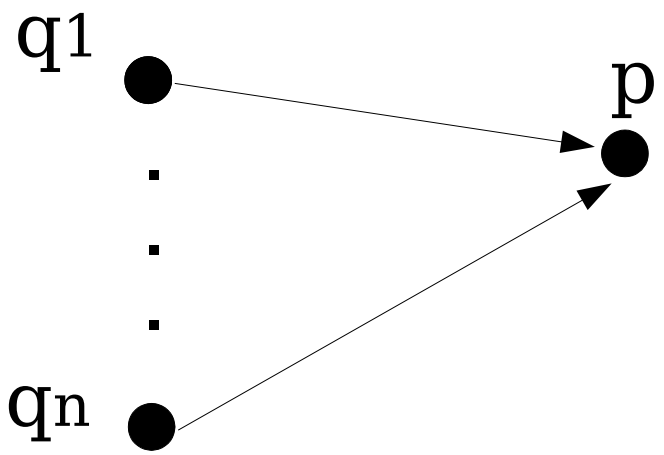


Authority と Hub の重み



Authority と Hub の重みのつけ方

Authority weight:x_p
Hub weight:y_p
($p \in$ base set)

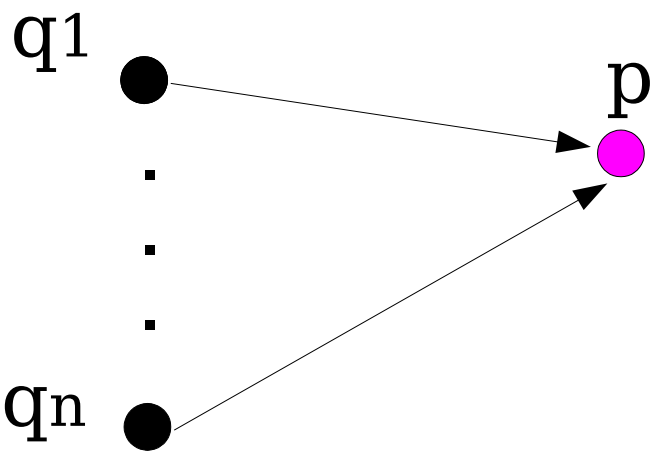


Authority と Hub の重みのつけ方

Authority weight:x_p

Hub weight:y_p

($p \in$ base set)



x_p := sum of y_q , for all q pointing to p

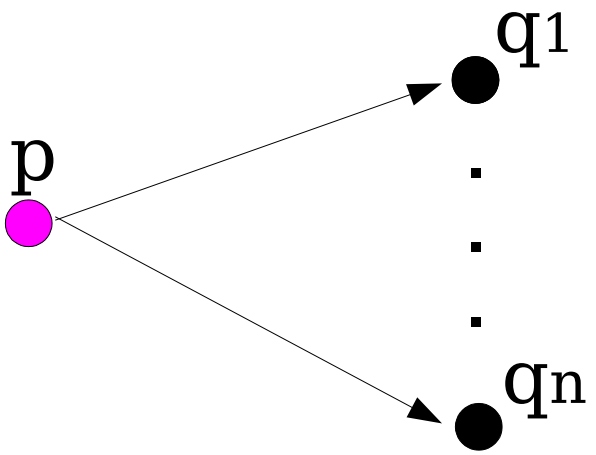
$$\longrightarrow \quad x_p = \sum_{q \ s.t \ q \to p} y_q$$

Authority と Hub の重みのつけ方

Authority weight:x_p

Hub weight:y_p

($p \in$ base set)



y_p := sum of x_q , for all q pointed to p

$$\longrightarrow \quad y_p = \sum_{p \ s.t \ p \to q} x_q$$

Authority と Hub の重みのつけ方

Authority weight：

$$x_p = \sum_{q \ s.t \ q \to p} y_q$$

Hub weight：

$$y_p = \sum_{p \ s.t \ p \to q} x_q$$

上記の2式を反復計算し
Authority weight と Hub weight を抽出

反復計算

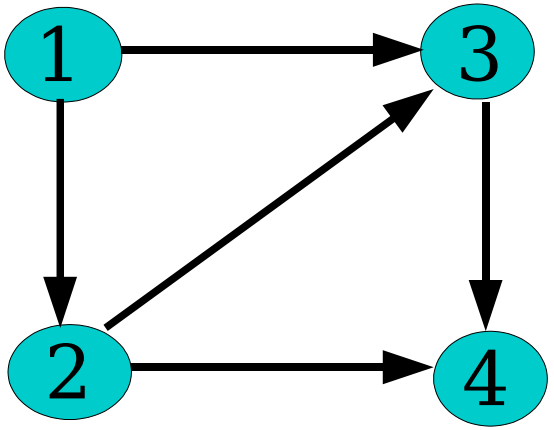
隣接行列:**A**

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \quad a_{ij} = \begin{cases} 1 & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$

反復計算

隣接行列:**A**

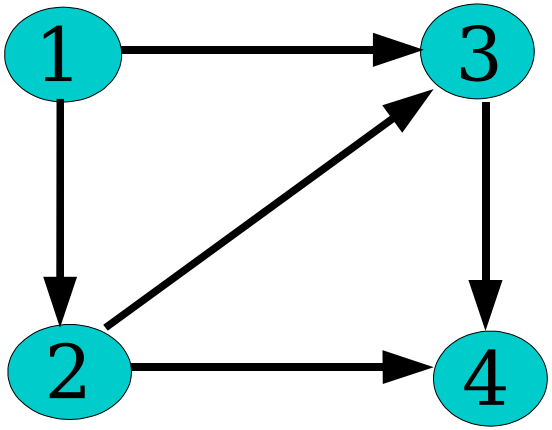
$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \quad a_{ij} = \begin{cases} 1 & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$



反復計算

隣接行列:**A**

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \qquad a_{ij} = \begin{cases} 1 & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$



$$\text{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

反復計算

A：隣接行列

$X_i = (x_1, x_2, \cdots, x_n)$

$Y_i = (y_1, y_2, \cdots, y_n)$

$Z = (1, 1, \cdots, 1) \in R^n$

k：反復回数

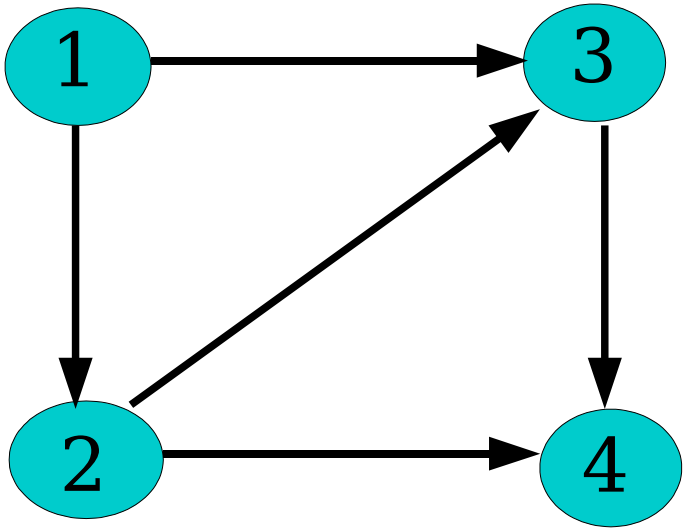
$(i = 1, 2, \cdots, k)$

初期値： $X_0 = Z$, $Y_0 = Z$

更新ルール： $X_i = A^T Y_{i-1}$

$Y_i = A X_i$

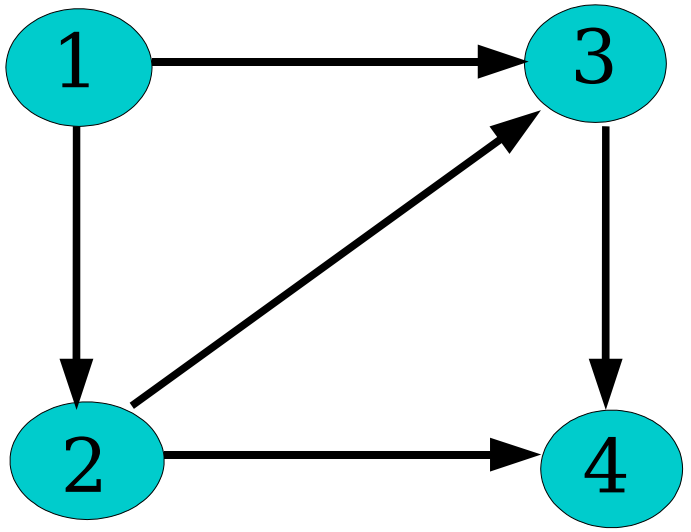
反復計算



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

反復計算

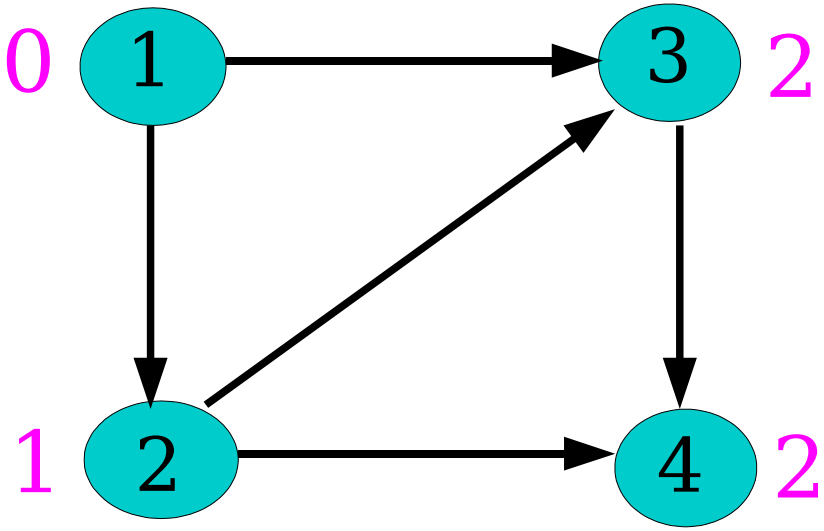
Authority weight $X_1 = A^T Y_0$



$$A^T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

反復計算

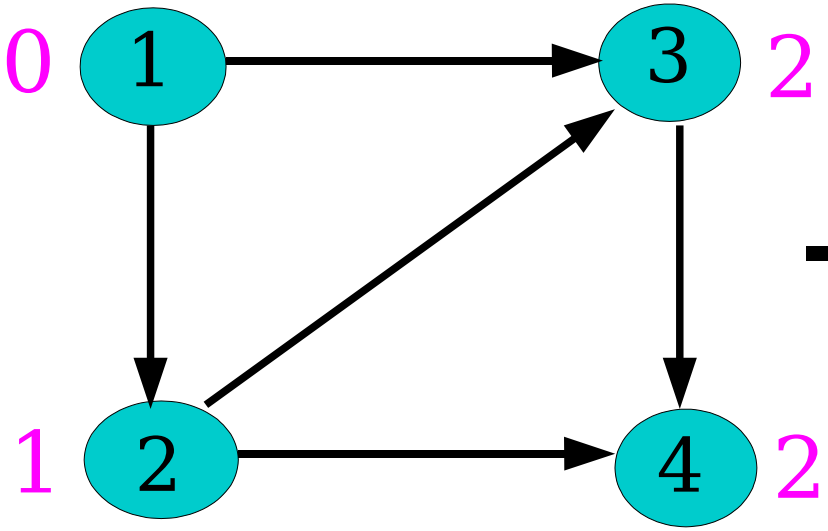
Authority weight $X_1 = A^T Y_0$



$$\begin{bmatrix} 0 \\ 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

反復計算

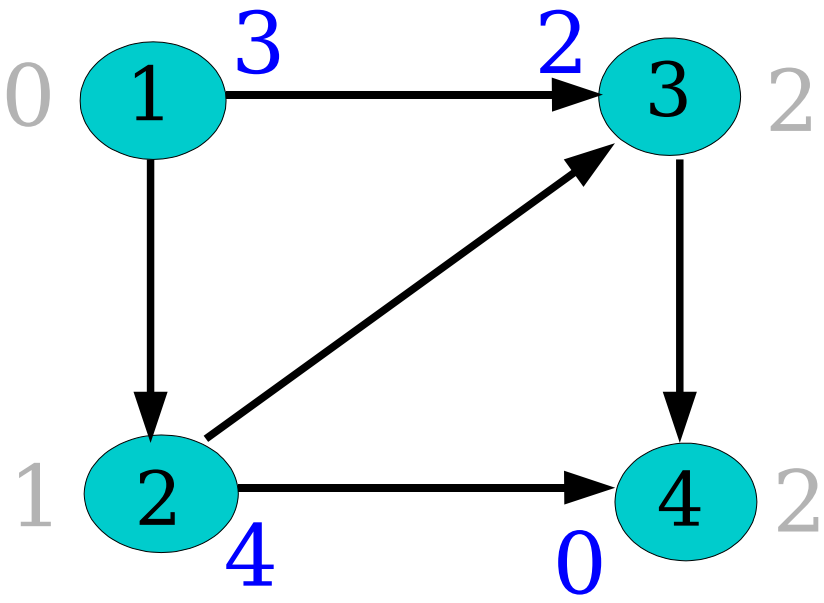
Hub weight $Y_1 = AX_1$



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

反復計算

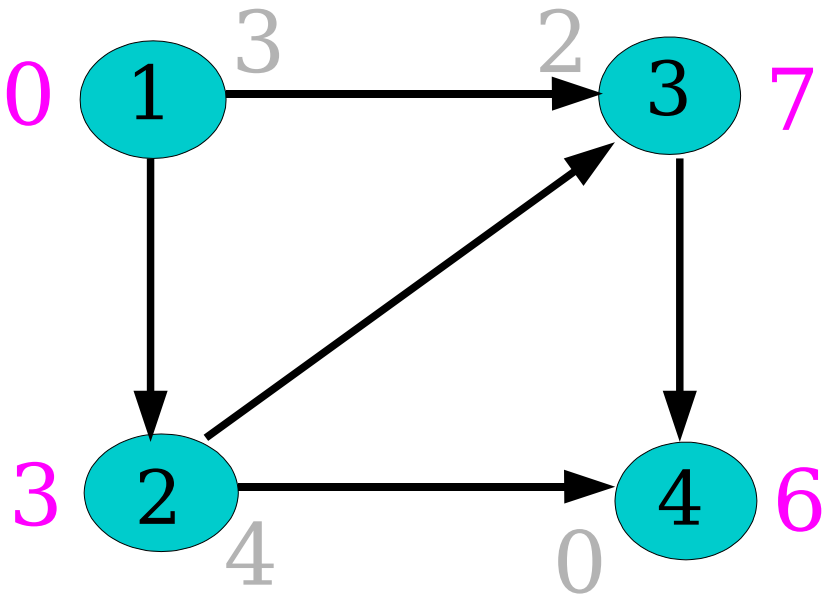
Hub weight $Y_1 = AX_1$



$$\begin{bmatrix} 3 \\ 4 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

反復計算

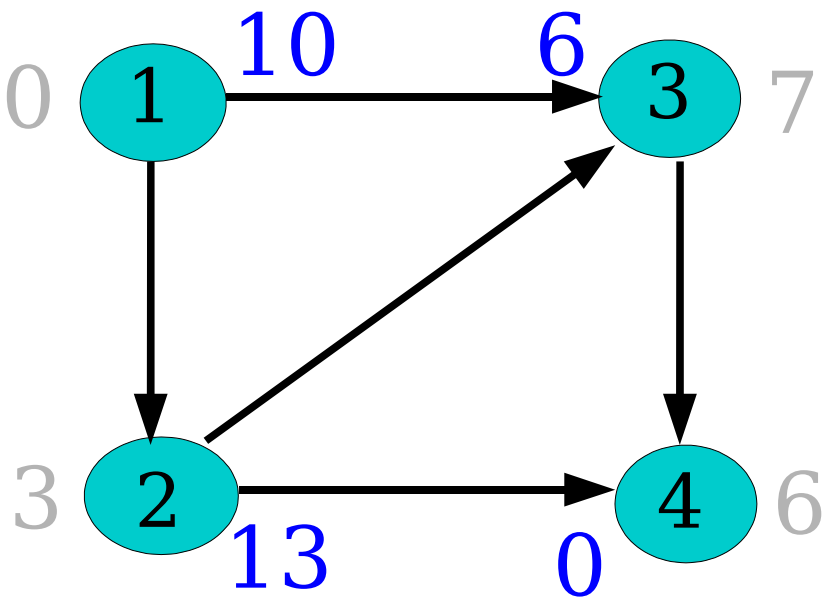
Authority weight $X_2 = A^T Y_1$



$$\begin{bmatrix} 0 \\ 3 \\ 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \\ 2 \\ 0 \end{bmatrix}$$

反復計算

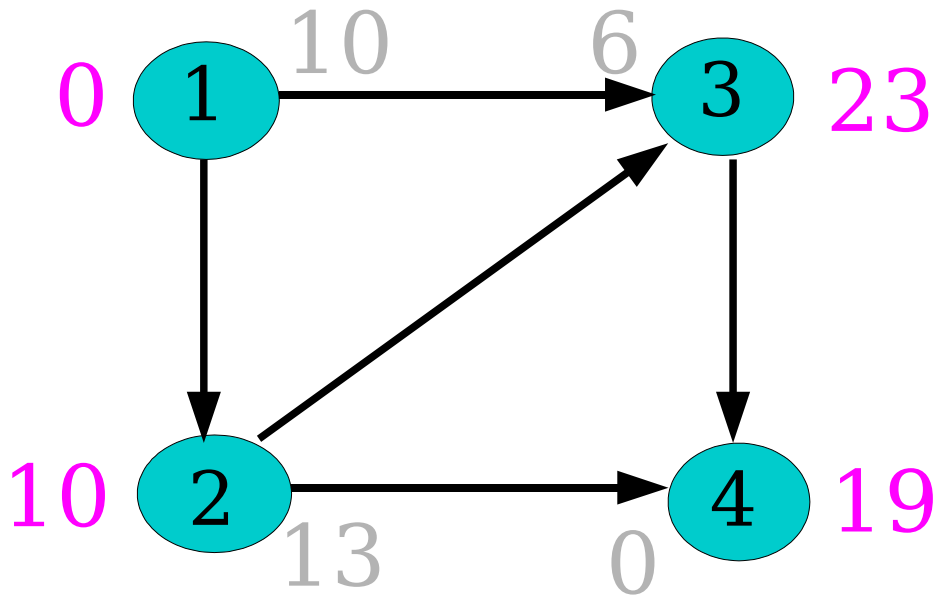
Hub weight $Y_2 = A X_2$



$$\begin{bmatrix} 10 \\ 13 \\ 6 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ 7 \\ 6 \end{bmatrix}$$

反復計算

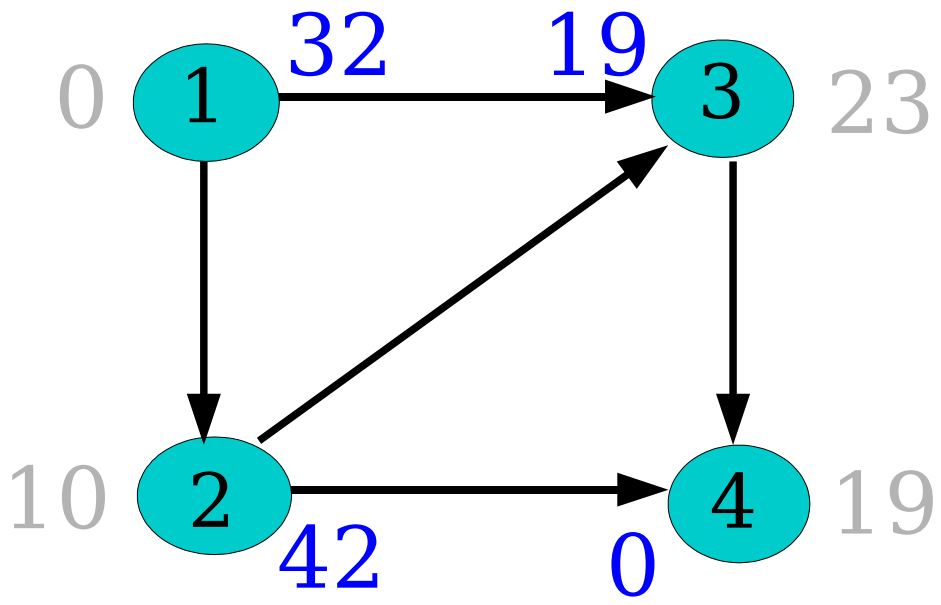
Authority weight $X_3 = A^T Y_2$



$$\begin{bmatrix} 0 \\ 10 \\ 23 \\ 19 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 10 \\ 13 \\ 6 \\ 0 \end{bmatrix}$$

反復計算

Hub weight $Y_3 = A X_3$



$$\begin{bmatrix} 32 \\ 42 \\ 19 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 23 \\ 19 \end{bmatrix}$$

目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

HITS アルゴリズムの改善点の概要

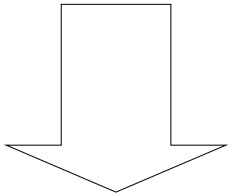
HITS アルゴリズムは各ページのコンテンツに
立ち入らない事が前提

HITS アルゴリズムの改善点の概要

HITS アルゴリズムは各ページのコンテンツに
立ち入らない事が前提 ⟶ topic drift 問題

HITS アルゴリズムの改善点の概要

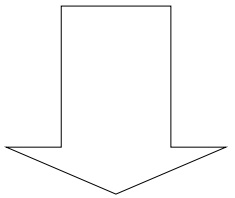
HITS アルゴリズムは各ページのコンテンツに
立ち入らない事が前提 → topic drift 問題



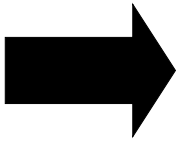
探したいトピックがページ内に含まれているならば
ページの weight を重くしてもよいのではないか？

HITS アルゴリズムの改善点の概要

HITS アルゴリズムは各ページのコンテンツに
立ち入らない事が前提 → topic drift 問題



探したいトピックがページ内に含まれているならば
ページの weight を重くしてもよいのではないか？



Web ページが半構造データであること
を利用

HITS アルゴリズムの改善点の概要

半構造データとは

“無構造データ”と“構造データ”の間のデータ構造のこと

HITS アルゴリズムの改善点の概要

半構造データとは

“無構造データ”と“構造データ”の間のデータ構造のこと

無構造データ：スキーマ(データベースで論理構造や物理構造を定めた仕様)が全く存在しないデータ
[ex. 普通の文章]

HITS アルゴリズムの改善点の概要

半構造データとは

“無構造データ”と“構造データ”の間のデータ構造のこと

無構造データ:スキーマ(データベースで論理構造や物理構造を定めた仕様)が全く存在しないデータ
[ex. 普通の文章]

構造データ:スキーマが決まっているデータ
[ex. RDB]

HITS アルゴリズムの改善点の概要

半構造データとは

“タグ”によって部分的に構造を記述できるデータ
[ex. XML や HTML]

無構造データ:スキーマ(データベースで論理構造や物理構造を定めた仕様)が全く存在しないデータ
[ex. 普通の文章]

構造データ:スキーマが決まっているデータ
[ex. RDB]

目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

tag weight の 導入

探したいトピックが特定の HTML タグで
強調されているならばその頻度を隣接行列に
反映させる

特定の HTML タグ

- <TITLE>～</TITLE>
- <STRONG>～</STRONG>
- <H1>～</H1>, . . . , <H6>～</H6> . . . 等

目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

anchor weight の 導入

探したいトピックが
「アンカーテキスト」「アンカータグ」
「アンカータグ前後のテキスト」
に含まれているならばその頻度を隣接行列に
反映させる

. . . . <A HREF =”http:// ...”>～</A>. . . .

anchor weight の 導入

探したいトピックが
「アンカーテキスト」「アンカータグ」
「アンカータグ前後のテキスト」
に含まれているならばその頻度を隣接行列に
反映させる

. . . . <A HREF =”http:// ...”>～</A>. . . .

トピック値を用いて隣接行列の改善を行う

トピック値の決定･隣接行列の改善

ページ p におけるトピック w の頻度を wp とし
トピックベクトル **W** を作成

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \qquad w_p = \text{ページ } p \text{ におけるトピック値}$$

隣接行列:**A**

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \qquad a_{ij} = \begin{cases} 1 + w_j & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$

トピック値の決定･隣接行列の改善

ページ p におけるトピック w の頻度を wp とし
トピックベクトル **W** を作成

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \qquad w_p = \text{ページ } p \text{ におけるトピック値}$$

隣接行列:**A**

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \qquad a_{ij} = \begin{cases} 1 + w_j & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$

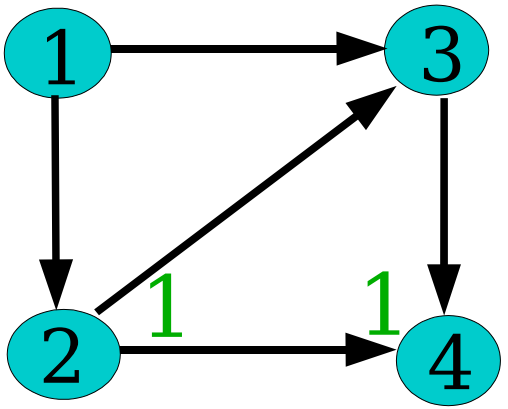
以降は HITS アルゴリズムと同様の手順で行う

トピック値を使用した反復計算

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$a_{ij} = \begin{cases} 1 + w_j & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$

$$W = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

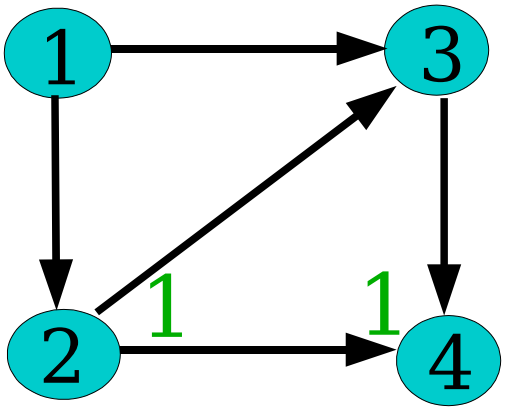


トピック値を使用した反復計算

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$a_{ij} = \begin{cases} 1 + w_j & \text{if page } i \text{ pointing to } j \\ 0 & \text{otherwise} \end{cases}$$

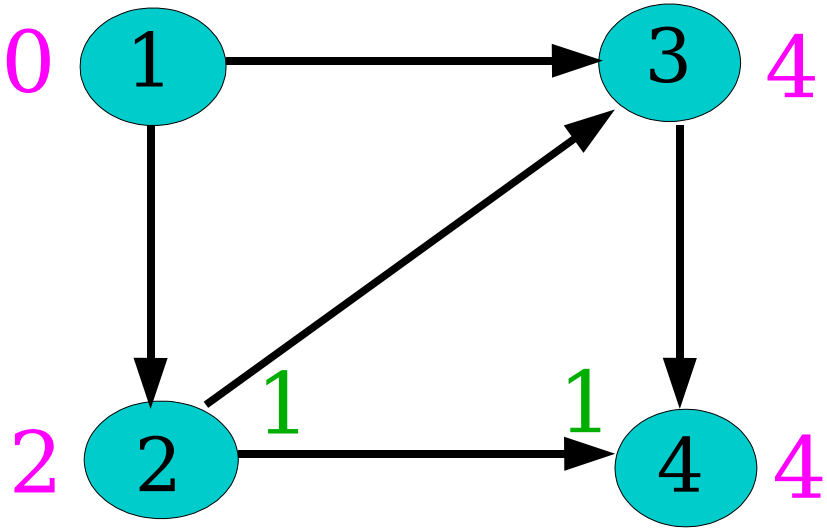
$$W = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$



$$A = \begin{bmatrix} 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

トピック値を使用した反復計算

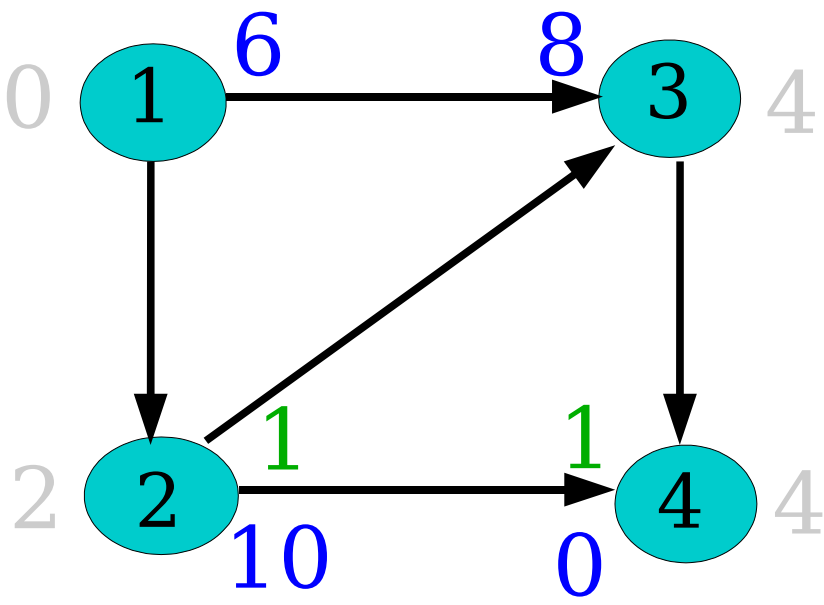
Authority weight $X_1 = A^T Y_0$



$$\begin{bmatrix} 0 \\ 2 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

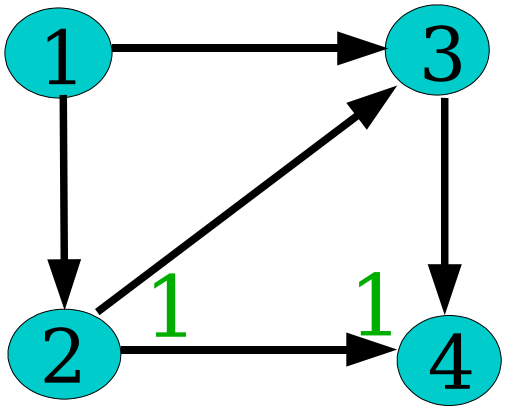
トピック値を使用した反復計算

Hub weight $Y_1 = AX_1$



$$\begin{bmatrix} 6 \\ 10 \\ 8 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ 4 \\ 4 \end{bmatrix}$$

トピック値を使用した反復計算



$$W = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \rightarrow \mathrm{A} = \begin{bmatrix} 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 2 \\ 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 6 \\ 10 \\ 8 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 12 \\ 16 \\ 26 \end{bmatrix}, \begin{bmatrix} 40 \\ 68 \\ 52 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 80 \\ 108 \\ 240 \end{bmatrix}, \begin{bmatrix} 268 \\ 582 \\ 480 \\ 0 \end{bmatrix} \cdots$$

X_0 Y_0 X_1 Y_1 X_2 Y_2 X_3 Y_3

目次

1. 研究の目的
2. HITS アルゴリズムの紹介
3. HITS アルゴリズムの改善
   - tag weight の 導入
   - anchor weight の 導入
4. 実験･考察

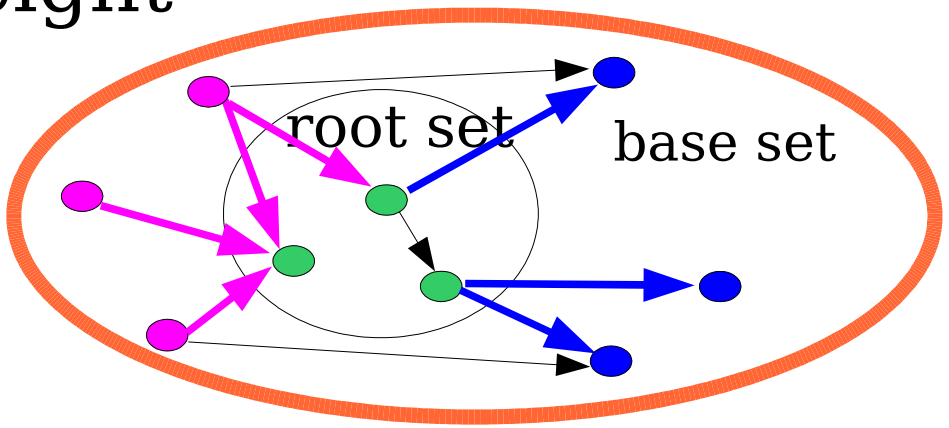
実験方法

比較対象

(a) Google サーチエンジン
(b) HITS アルゴリズム
(c) HITS + tag weight
(d) HITS + anchor weight

パラメータ

root set ：100
base set：1000～5000



root set の収集方法

Google Web API を利用

評価方法

HITS アルゴリズムの評価を行った
研究グループを参考：アンケート形式

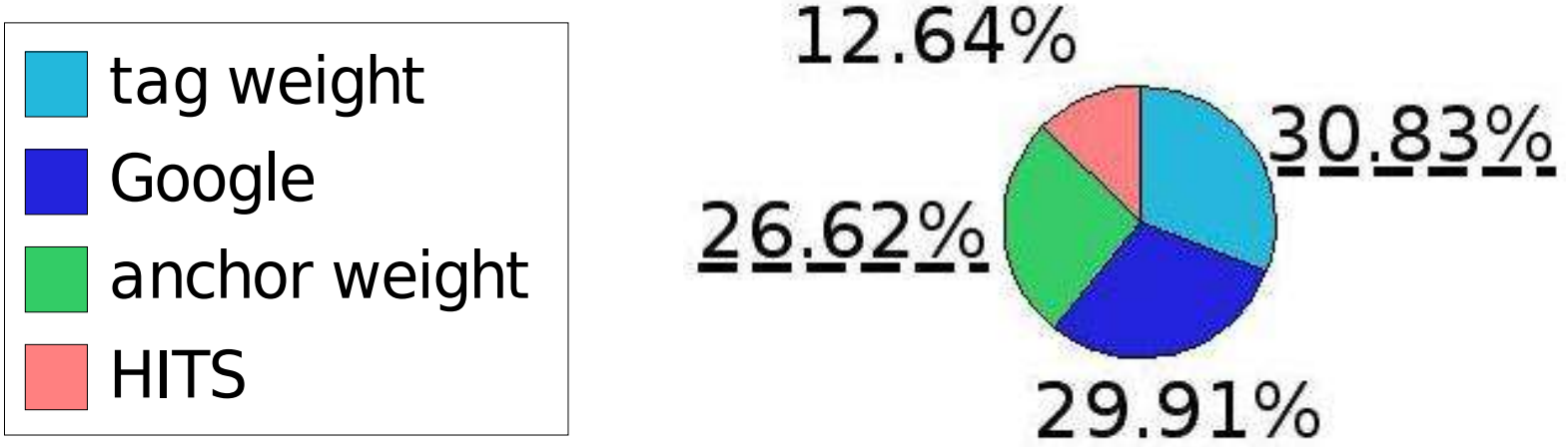
- 対象者
  → 谷研究室　20名

- アンケート方法
  → 各々の上位3つのWeb ページを閲覧し
  「good」「fair」「bad」の3段階で評価

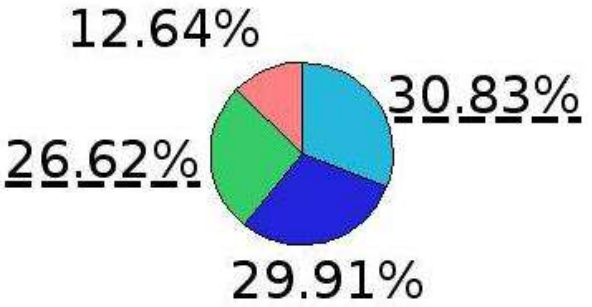
注:どの方法がどの Web ページを返したかは
伏せて行った

実験結果

- HITS + tag weight : 30.83%
- Google サーチエンジン : 29.91%
- HITS + anchor weight : 26.62%
- HITS アルゴリズム : 12.64%



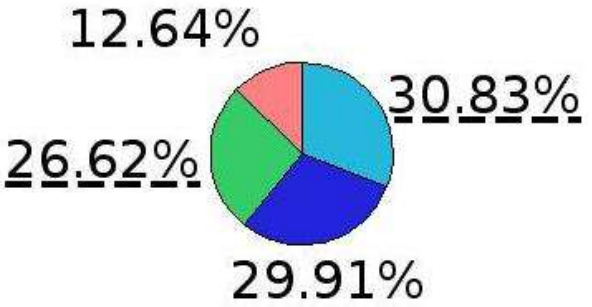
考察 1



- HITS + tag weight : 30.83%
- Google サーチエンジン : 29.91%
- HITS + anchor weight : 26.62%
- HITS アルゴリズム : 12.64%

→ 各 Web ページのコンテンツに立ち入りページを評価することは有効

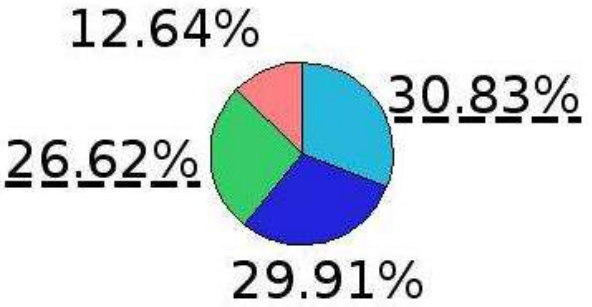
考察 2



- HITS + tag weight : 30.83%
- **Google** サーチエンジン : 29.91%
- HITS + anchor weight : 26.62%
- HITS アルゴリズム : 12.64%

→ トピックが固有名詞のとき
Google は正式なページが上位にくる

考察 3



- HITS + tag weight : 30.83%
- Google サーチエンジン : 29.91%
- HITS + anchor weight : 26.62%
- HITS アルゴリズム : 12.64%

→ どんなトピックを検索しても
Google は企業のページが上位にくる

今後の課題

＊ドメイン名に注目したページの重み付け
＊HTML タグの種類による重みの調節
＊複数のトピックによる検索

今後の課題

* ドメイン名に注目したページの重み付け
* HTML タグの種類による重みの調節
* 複数のトピックによる検索

→ URL にトピックが含まれ
ドメイン名が “co.jp” や “ac.jp” 等の場合
ページの weight を重くする

今後の課題

＊ドメイン名に注目したページの重み付け

＊HTML タグの種類による重みの調節

＊複数のトピックによる検索

→ tag weight 導入の際
<TITLE>～</TITLE> 等で
トピックが挟まれていた場合
ページの weight を重くする

今後の課題

＊ドメイン名に注目したページの重み付け
＊HTML タグの種類による重みの調節
＊複数のトピックによる検索

→ 今回の実験ではトピックを1つに限定したが
複数のトピックで検索することにより
また別の結果が得られるかも知れない

おわり