# title

author

## 1 Eigen Factor$^{\mathrm{TM}}$ の原理

Eigen Factor$^{\mathrm{TM}}$（以下，EF と省略する）は，論文誌をランキング評価するための定量的な指標である．他の評価指標として．Impact Factor（以下，IF と省略する）が知られている．EF は IF の短所を改善したものと位置づけられる．IF との比較については次の節で述べ，この節では EF の原理について述べる．

EF の原理は Google の PageRank の原理に類似している．論文誌の重要度を測る基準として，次の 3 つの基準が挙げられる．

1. よく引用される雑誌は重要度が高い．
2. 重要な雑誌から引用される雑誌もまた重要度が高い．
3. 多く引用する雑誌（例：レビュー誌）からの引用は重要度が低い．

これらの基準に照らして EF は計算される．

論文誌の総数は $N$ で，それぞれの論文誌を添字集合 $\{1, 2, \ldots, N\}$ によって識別する．論文誌 $i$ が論文誌 $j$ を引用している場合，$A_{ij} = 0$ $(i \neq j)$．そうでない場合，$A_{ij} = 0$ として隣接行列 $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ を定義する．ただし，$A_{ii} = 0$ とし，論文誌のセルフ引用はカウントしない．次に，行列 $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ を以下のように定める．

$$
\begin{aligned}
&B_{ij} = \\
&\text{（論文誌 } i \text{ から論文誌 } j \text{ への遷移確率）} = \\
&\begin{cases}
\alpha \dfrac{A_{ij}}{\sum_{\ell=1}^{N} A_{i\ell}} + (1-\alpha)\mathbf{1}\boldsymbol{\nu}^{\top} & \sum_{\ell=1}^{N} A_{i\ell} \geq 1 \text{ の場合.} \\
\mathbf{1}\boldsymbol{\nu}^{\top} & \sum_{\ell=1}^{N} A_{i\ell} = 0 \text{ の場合.}
\end{cases}
\end{aligned}
$$

ここで，$\boldsymbol{\nu} \in \mathbb{R}^{N}$ はパーソナル化ベクトルとよばれ，

$$
\nu_i = \frac{\text{EF の計測期間中に論文誌 } i \text{ が発行した論文の総数}}{\text{EF の計測期間中に発行された論文の総数}}
$$

で定義する．また，$\alpha$ は PageRank に則り，0.85 とすることが多い．

$$
\boldsymbol{\pi}^{\top} = \boldsymbol{\pi}^{\top}\boldsymbol{B} \text{ かつ } \mathbf{1}^{\top}\boldsymbol{\pi} = 1
$$

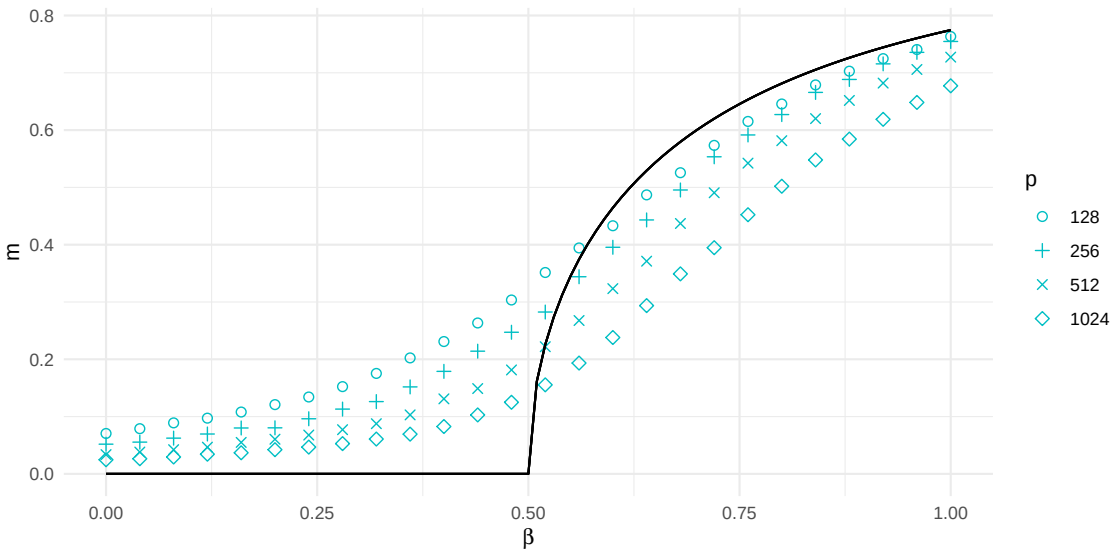


**図 1:** bbp 相転移

を満たす $\boldsymbol{\pi} \in \mathbb{R}^N$ がページランクとなる．ページランクを用いて EF を次のように計算する．

$$\text{論文誌 } i \text{ の EF} = 100 \times \frac{(\boldsymbol{\pi}^\top \boldsymbol{C})_i}{\sum_{\ell=1}^{N} (\boldsymbol{\pi}^\top \boldsymbol{C})_\ell}$$

ただし，行列 $\boldsymbol{C}$ は行列 $B$ からテレポーテーション項を除いた行列で，次で定義する [2]．

$$C_{ij} = \begin{cases} \dfrac{A_{ij}}{\sum_{\ell=1}^{N} A_{i\ell}} & \sum_{\ell=1}^{N} A_{i\ell} \geq 1 \text{ の場合.} \\ 0 & \sum_{\ell=1}^{N} A_{i\ell} = 0 \text{ の場合.} \end{cases}$$

直感的には，遷移行列 $\boldsymbol{B}$ の定常分布ベクトル $\boldsymbol{\pi}$ を，テレポーテーションを含まない遷移行列 $\boldsymbol{C}$ で更に 1 ステップ遷移させた状態分布に 100 をかけた数値が EF である．

## 2 Impact Factor との比較

IF は次のように計算される．

$$\text{論文誌 } i \text{ の IF} = \frac{\sum_{\ell=1}^{N} A_{\ell i}}{\text{IF の計測期間中に論文誌 } i \text{ が発行した論文の総数}}.$$

IF と比較したときの EF の長所短所について述べる．

### 2.1 EF の長所 [1]

- 基準 1 だけでなく，基準 2,3 も考慮している点．
  計算式からわかるように，よく引用される論文誌ほど分子の項が大きくなり，IF の値は大きくなる．これは基準 1 を反映している．一方，EF は PageRank と同様の性質をもつため，基準 2,3 にも則る．これにより，EF は単なる被引用数だけでなく，どの論文誌から引用されたかまで考慮することができている．
- EF の計測期間は IF の計測期間よりも長い点．
  よく使われる IF の計測期間が 2 年であるのに対し，EF の計測期間は 5 年と IF の計測期間よりも長い．計測期間が長いほど，引用数も増加するため，より広範な評価が可能となる．
- 論文誌のセルフ引用を無視する点．
  EF では隣接行列 $\boldsymbol{A}$ の $A_{ii}$ 成分を 0 とするが，IF ではとくにそのような操作は行わない．$A_{ii} = 0$ とすることで，IF を高めたい論文誌が自誌の論文をセルフ引用するような小技を阻止している．

### 2.2 EF の短所 [2]

- 総被引用数と相関がある点．
  EF も IF と同様に基準 1 を考慮しているため，単に総被引用数を増やせば EF のスコアが高くなる傾向がある．EF はセルフ引用は除外するが，同出版社の別タイトル論文誌が論文誌間のセルフ引用のような手の込んだ小技を阻止できない．

## 3 EF の改善法の提案

さきほど述べたように，EF にはいくつかの短所がある．この節ではその短所を改善する方法について考察する．

EF が抱える，同出版社の別タイトル論文誌が論文誌間のセルフ引用のような手の込んだ小技を阻止できない，という問題に注目する．これはウェブサイトの検索エンジン最適化（SEO）におけるリンクファームに類似している．リンクファームとは，ウェブサイト群が相互リンクすることでウェブサイト群全体のページランクを故意に増加させる，SEO スパムの一種である．EF は PageRank のアルゴリズムとほぼ同様であるため，PageRank のリンクファーム対策がそのまま EF の問題点に適用できると考えられる．PageRank のリンクファーム対策として次の手法が提案されている [4]．以下，EF の文脈に即して， [4] のアルゴリズムの概要を述べる．

- **SeedSet の検出**
  1. 集合 $In(i), Out(i)$ $i \in \{1, \ldots, N\}$ を空集合で初期化する．

2. 論文誌 $i$ を引用する論文誌の集合を $In(i)$ とする．
3. 論文誌 $i$ が引用する論文誌の集合を $Out(i)$ とする．
4. $In(i) \cap Out(i)$ の要素数がしきい値以上の場合，$i$ をリンクファームの疑いがある論文誌として登録する．
5. 全ての $i \in \{1, \ldots, N\}$ について以上を行う．
6. 疑わしいとされた論文誌の集合を SeedSet とする．

- **SeedSet の拡張**
  1. 論文誌 $i$ が引用する論文誌の集合を $Out(i)$ とする．
  2. $Out(i) \cap$ SeedSet の要素数がしきい値以上なら論文誌 $i$ を SeedSet に追加する．
  3. 全ての $i \in \{1, \ldots, N\}$ について以上を繰り返す．SeedSet が変化しなくなったら終了．
- **リンクファームへのペナルティ**

  論文誌 $i$ が SeedSet に属する論文誌 $j$ を引用している場合，
  - $A_{ij} = 0$ とする．

    または
  - 論文誌 $i$ が SeedSet に属する論文誌を $n$ 冊引用している場合，$A_{ij} = 1/n$ とする．

「SeedSet の検出」はリンクファームを形成する論文誌群は相互に引用することを利用している．「SeedSet の拡張」は SeedSet に属する論文誌を引用する論文誌も疑わしい可能性があることを利用している．最終的に得られた SeedSet がリンクファームに属する論文誌の集合を表す．「リンクファームへのペナルティ」については，リンクファームに属する論文誌の EF を直接下げる，などこの他の方法も考えられる．

実際にこの手法を試してみることは行っていない．ただ，EF と PageRank アルゴリズムの類似性により，SEO スパム対策のアルゴリズムが論文誌の評価において現れる種々の問題に対して有効であると期待できそうである．EF の他の問題点についてもこのようなアナロジーが有効かどうかは引き続き調査していきたい．

## 4 Eigen Factor$^{\mathrm{TM}}$ の原理

Eigen Factor$^{\mathrm{TM}}$（以下，EF と省略する）は，論文誌をランキング評価するための定量的な指標である．他の評価指標として．Impact Factor（以下，IF と省略する）が知られている．EF は IF の短所を改善したものと位置づけられる．IF との比較については次の節で述べ，この節では EF の原理について述べる．

EF の原理は Google の PageRank の原理に類似している．論文誌の重要度を測る基準として，次の 3 つの基準が挙げられる．

1. よく引用される雑誌は重要度が高い．
2. 重要な雑誌から引用される雑誌もまた重要度が高い．
3. 多く引用する雑誌（例：レビュー誌）からの引用は重要度が低い．

これらの基準に照らして EF は計算される．

論文誌の総数は $N$ で，それぞれの論文誌を添字集合 $\{1, 2, \ldots, N\}$ によって識別する．論文誌 $i$ が論文誌 $j$ を引用している場合，$A_{ij} = 0$ $(i \neq j)$．そうでない場合，$A_{ij} = 0$ として隣接行列 $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ を定義する．ただし，$A_{ii} = 0$ とし，論文誌のセルフ引用はカウントしない．次に，行列 $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ を以下のように定める．

$$B_{ij} = \tag{1}$$

$$(\text{論文誌 } i \text{ から論文誌 } j \text{ への遷移確率}) = \tag{2}$$

$$\begin{cases} \alpha \frac{A_{ij}}{\sum_{\ell=1}^{N} A_{i\ell}} + (1-\alpha)\mathbf{1}\boldsymbol{\nu}^{\top} & \sum_{\ell=1}^{N} A_{i\ell} \geq 1 \text{ の場合.} \\ \mathbf{1}\boldsymbol{\nu}^{\top} & \sum_{\ell=1}^{N} A_{i\ell} = 0 \text{ の場合.} \end{cases} \tag{3}$$

ここで，$\boldsymbol{\nu} \in \mathbb{R}^{N}$ はパーソナル化ベクトルとよばれ，

$$\nu_i = \frac{\text{EF の計測期間中に論文誌 } i \text{ が発行した論文の総数}}{\text{EF の計測期間中に発行された論文の総数}} \tag{4}$$

で定義する．また，$\alpha$ は PageRank に則り，0.85 とすることが多い．

$$\boldsymbol{\pi}^{\top} = \boldsymbol{\pi}^{\top}\boldsymbol{B} \text{ かつ } \mathbf{1}^{\top}\boldsymbol{\pi} = 1 \tag{5}$$

を満たす $\boldsymbol{\pi} \in \mathbb{R}^N$ がページランクとなる．ページランクを用いて EF を次のように計算する．

$$\text{論文誌 } i \text{ の EF} = 100 \times \frac{(\boldsymbol{\pi}^\top \boldsymbol{C})_i}{\sum_{\ell=1}^{N} (\boldsymbol{\pi}^\top \boldsymbol{C})_\ell} \tag{6}$$

ただし，行列 $\boldsymbol{C}$ は行列 $B$ からテレポーテーション項を除いた行列で，次で定義する [2]．

$$C_{ij} = \begin{cases} \dfrac{A_{ij}}{\sum_{\ell=1}^{N} A_{i\ell}} & \sum_{\ell=1}^{N} A_{i\ell} \geq 1 \text{ の場合.} \\ 0 & \sum_{\ell=1}^{N} A_{i\ell} = 0 \text{ の場合.} \end{cases} \tag{7}$$

直感的には，遷移行列 $\boldsymbol{B}$ の定常分布ベクトル $\boldsymbol{\pi}$ を，テレポーテーションを含まない遷移行列 $\boldsymbol{C}$ で更に 1 ステップ遷移させた状態分布に 100 をかけた数値が EF である．

## 5 Impact Factor との比較

IF は次のように計算される．

$$\text{論文誌 } i \text{ の IF} = \tag{8}$$

$$\frac{\sum_{\ell=1}^{N} A_{\ell i}}{\text{IF の計測期間中に論文誌 } i \text{ が発行した論文の総数}}. \tag{9}$$

IF と比較したときの EF の長所短所について述べる．

### 5.1 EF の長所 [1]

- 基準 1 だけでなく，基準 2,3 も考慮している点．
  計算式からわかるように，よく引用される論文誌ほど分子の項が大きくなり，IF の値は大きくなる．これは基準 1 を反映している．一方，EF は PageRank と同様の性質をもつため，基準 2,3 にも則る．これにより，EF は単なる被引用数だけでなく，どの論文誌から引用されたかまで考慮することができている．
- EF の計測期間は IF の計測期間よりも長い点．
  よく使われる IF の計測期間が 2 年であるのに対し，EF の計測期間は 5 年と IF の計測期間よりも長い．計測期間が長いほど，引用数も増加するため，より広範な評価が可能となる．
- 論文誌のセルフ引用を無視する点．
  EF では隣接行列 $\boldsymbol{A}$ の $A_{ii}$ 成分を 0 とするが，IF ではとくにそのような操作は行わない．$A_{ii} = 0$ とすることで，IF を高めたい論文誌が自誌の論文をセルフ引用するような小技を阻止している．

### 5.2 EF の短所 [2]

- 総被引用数と相関がある点．
  EF も IF と同様に基準 1 を考慮しているため，単に総被引用数を増やせば EF のスコアが高くなる傾向がある．EF はセルフ引用は除外するが，同出版社の別タイトル論文誌が論文誌間のセルフ引用のような手の込んだ小技を阻止できない．

## 6 EF の改善法の提案

さきほど述べたように，EF にはいくつかの短所がある．この節ではその短所を改善する方法について考察する．

EF が抱える，同出版社の別タイトル論文誌が論文誌間のセルフ引用のような手の込んだ小技を阻止できない，という問題に注目する．これはウェブサイトの検索エンジン最適化（SEO）におけるリンクファームに類似している．リンクファームとは，ウェブサイト群が相互リンクすることでウェブサイト群全体のページランクを故意に増加させる，SEO スパムの一種である．EF は PageRank のアルゴリズムとほぼ同様であるため，PageRank のリンクファーム対策がそのまま EF の問題点に適用できると考えられる．PageRank のリンクファーム対策として次の手法が提案されている [4]．以下，EF の文脈に即して， [4] のアルゴリズムの概要を述べる．

- **SeedSet の検出**
  1. 集合 $In(i), Out(i)$ $i \in \{1, \ldots, N\}$ を空集合で初期化する．

2. 論文誌 $i$ を引用する論文誌の集合を $In(i)$ とする．
3. 論文誌 $i$ が引用する論文誌の集合を $Out(i)$ とする．
4. $In(i) \cap Out(i)$ の要素数がしきい値以上の場合，$i$ をリンクファームの疑いがある論文誌として登録する．
5. 全ての $i \in \{1, \ldots, N\}$ について以上を行う．
6. 疑わしいとされた論文誌の集合を SeedSet とする．

- **SeedSet の拡張**
  1. 論文誌 $i$ が引用する論文誌の集合を $Out(i)$ とする．
  2. $Out(i) \cap$ SeedSet の要素数がしきい値以上なら論文誌 $i$ を SeedSet に追加する．
  3. 全ての $i \in \{1, \ldots, N\}$ について以上を繰り返す．SeedSet が変化しなくなったら終了．
- リンクファームへのペナルティ

  論文誌 $i$ が SeedSet に属する論文誌 $j$ を引用している場合，
  - $A_{ij} = 0$ とする．

    または
  - 論文誌 $i$ が SeedSet に属する論文誌を $n$ 冊引用している場合，$A_{ij} = 1/n$ とする．

「SeedSet の検出」はリンクファームを形成する論文誌群は相互に引用することを利用している．「SeedSet の拡張」は SeedSet に属する論文誌を引用する論文誌も疑わしい可能性があることを利用している．最終的に得られた SeedSet がリンクファームに属する論文誌の集合を表す．「リンクファームへのペナルティ」については，リンクファームに属する論文誌の EF を直接下げる，などこの他の方法も考えられる．

実際にこの手法を試してみることは行っていない．ただ，EF と PageRank アルゴリズムの類似性により，SEO スパム対策のアルゴリズムが論文誌の評価において現れる種々の問題に対して有効であると期待できそうである．EF の他の問題点についてもこのようなアナロジーが有効かどうかは引き続き調査していきたい． [3]

## 7

### 7.1

## 参考文献

[1]M. Franceschet, “Ten good reasons to use the eigenfactortm metrics”, Inf. Process. Manage. **46**, 555–558 (2010).

[2]増田 直紀, “アイゲンファクターを知る”, 統計数理 **61**, 147–166 (2013).

[3]NASA, *Beginner’s Guide to Kites*, https://www.grc.nasa.gov/www/k-12/airplane/shortk.html, Accessed on 2020-01-08.

[4]B. Wu and B. D. Davison, “Identifying link farm spam pages”, in Special interest tracks and posters of the 14th international conference on world wide web (2005), pp. 820–829.