

ウェブページのランク付け指標の妥当性の検証

東京情報大・総合情報 ○永田 裕史, 今井 哲郎, 吉澤 康介, 三宅 修平

1. はじめに

インターネット上のウェブページは、ハイパーリンクと呼ばれる有向リンクで接続され、巨大なグラフ構造を形成している。ウェブ検索システムは、このグラフ構造で表現されるウェブページを機械的にランク付けし、ユーザーにとって有益なウェブページを提示する際の一つの判断材料としている。ウェブページの優先順位付けを行うための著名な指標として、PageRankとHITSアルゴリズムが知られている。

本稿では、Wikipediaのページ群の一部をPageRankとHITSアルゴリズムの2つの指標で評価し、それらを実際のページビューと比較することで、ウェブページの優先順位付けを行うための指標としての問題点を明らかにする。

2. 分析方法

2.1. 分析対象

本稿では、分析対象としてWikipediaのページ群を取り上げる。本来であれば、インターネット上のウェブページ全体を対象として分析を行うべきである。しかしウェブページ数は膨大であり、ある時点でのダンプデータの取得も困難で、またウェブページの閲覧回数(ページビュー)の取得も難しいといった問題がある。そのため、インターネットに存在するウェブページ全てを分析対象とすることは、現実的ではない。Wikipediaには、41,190,504の記事が存在し(2016年9月1日現在)、定期的にウェブページの情報がダンプデータとして保管・公開されている(*1)。またページビューの情報も比較的容易に取得可能である(*2)。このことから本稿では、インターネット上のウェブページ全体のサンプルとしてWikipediaのページ群を取り上げ、分析対象として設定した。

今回用いたWikipediaのデータは、Wikipediaの内部のハイパーリンクで構成されるグラフ構造のデータと、ページビューのデータである。グラフ構造のデータは2016年7月

1日時点のダンプデータから作成されたものであり、ページビューのデータは2016年3月～2016年6月のページビューを合算して作成されたものである。

2.2. ウェブページのランク付け指標

ウェブページの重要度を機械的にランク付けするための方法として、本稿ではPageRankとHITSを取り上げる。いずれの指標も、ウェブページをノード、ウェブページ間を結ぶハイパーリンクを有向リンクとするグラフ構造のみを利用して、ウェブページの重要度を算出する。

2.2.1. PageRank

PageRankはGoogleの創設者であるBrinとPageによって1998年に発表された[1]もので、Googleの検索システムにも利用されたことで知られている。あるページXのPageRank値 $PR(X)$ は、以下の式で与えられる。

$$PR(X) = \frac{(1-d)}{n} + d \sum_{X' \in N_{in}(X)} \frac{PR(X')}{|N_{out}(X')|}$$

ここで d はランダムジャンプをしない確率($0 \leq d < 1$, 通常は0.85程度)、 n はノードの総数、 $N_{in}(X)$ はノードXへのリンクを持つノードの集合、 $|N_{out}(X)|$ はノードXの出次数を表す。

あるウェブページのPageRank値は、ランダムサーファークがランダムにリンクを辿る遷移を十分な回数行った後に、そのウェブサイトに存在する確率に等しい。

2.2.2. HITS

HITSアルゴリズムでは、全てのノードにハブスコアとオーソリティスコアの2種類のスコアを割り当てる。あるウェブサイトXのハブスコア $H(X)$ とオーソリティスコア $A(X)$ は、以下の式を満たすように定義される。

$$A(X) = \sum_{X' \in N_{in}(X)} H(X')$$

$$H(X) = \sum_{X' \in N_{in}(X)} A(X')$$

Validity verification of ranking indexes of web pages

Hiroshi NAGATA, Tetsuo IMAI, Kousuke YOSHIZAWA and Shuhei MIYAKE



図 1 カテゴリー「複雑系」に属する Wikipedia の記事が構成するネットワークの構造

ここで $N_{in}(X)$ はノードXへのリンクを持つノードの集合を表す．なお $H(X)$ と $A(X)$ はそれぞれ，全てのノードXに関する総和が1となるように正規化される．

ハブスコアの高いウェブページは，質の高いウェブページに多数のリンクを張っているウェブページであり，オーソリティスコアの高いウェブページは，質の高いウェブページから多数のリンクを張られているウェブページである．

3. 分析結果と考察

カテゴリー「複雑系」に属する Wikipedia のページについて可視化したものを，図 1 に示す．図 1 における各ノードは Wikipedia の各記事を表しており，ページビューが多い記事は

ど大きなノードとして描かれている．また各ノードの色は，PageRank 値の小→中→大に応じて青→赤→黄と塗り分けられている．各ノード下部の棒グラフは，左からページビューの順位，PageRank の順位，オーソリティスコアの順位，ハブスコアの順位を示し，高い順位が高い棒で示されている．記事の数を表す総ノード数は160，記事間の内部リンクの数を表す総リンク数は2,473であった．

ページビュー，PageRank，オーソリティスコア，ハブスコアの間に関連行列を，表 1 に示す．ページビューと PageRank の間にはやや高い相関があった(0.549)が，HITS のオーソリティスコア/ハブスコアとページビューの間はほぼ無相関(-0.165/-0.171)であった．またオーソリティスコアとハブスコアとの間には非

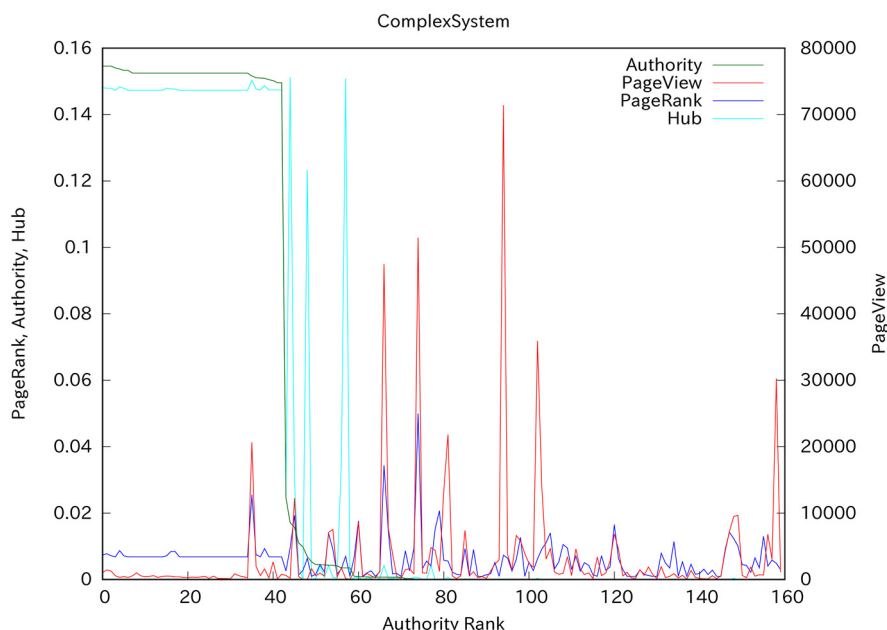


図 2 オーソリティスコアの順位に対する各指標の値

常に強い相関(0.961)が見られた。なお、カテゴリー「日本の科学者」(総ノード数 6,359, 総リンク数 12,807)に関しても調査をしたが(表 2), こちらについても同様の結果となった。

以下では, HITS の 2 つの指標がページビューを反映できていない原因を考える。

図 2 は, 各指標の値をプロットしたグラフである。横軸はオーソリティスコアの順位を, 縦軸は各指標の値を指す。図 2 から, 非常に高いオーソリティスコアとハブスコアは, 1~45 位前後までのノードで占められていることが分かる。具体的な記事を見てみると, オーソリティスコアの 1~46 位までの全ての記事が, 「ライフゲーム」に関する記事であった。これらは図 1 において, 左上に示された密に結合した部分である。これらの記事は密な相互リンクを形成し, コミュニティを形成していることが分かる。HITS の 2 つの指標はこのコミュニティに引

表 1 カテゴリー「複雑系」に属する記事に関する, ページビューと各ランク付け指標との相関行列

	PV	PR	Auth.	Hub
PageView	1.000	0.549	-0.165	-0.171
PageRank		1.000	0.131	0.145
Authority			1.000	0.961
Hub				1.000

張られる形で配点してしまっているために, う

まくページビューを反映できていないものと考えられる。

以上の結果から, 既存のウェブページのランク付け指標は, 必ずしもページビューをうまく反映できていないことが分かった。特に HITS に関しては, 密な部分グラフに対する頑健性が課題であると考えられる。

表 2 カテゴリー「日本の科学者」に属する記事に関する, ページビューと各ランク付け指標との相関行列

	PV	PR	Auth.	Hub
PageView	1.000	0.326	0.046	0.053
PageRank		1.000	0.369	0.378
Authority			1.000	0.987
Hub				1.000

4. おわりに

本稿では, ウェブページのランク付け指標の妥当性を検証するために, ウェブページのサンプルとして Wikipedia のページ群を取り上げ, PageRank と HITS アルゴリズムの 2 つの指標と, 実際のページビューがどの程度の相関を持つのかを比較した。その結果, PageRank と HITS の 2 指標だけでは実際のページビューの高さを測ることが難しいこと, また特に HITS の 2 つの指標が, 密な相互リンクを有する部分グラフによって強い影響を受けやすいことを

明らかにした。今後の課題として、より大きなサンプルについて分析をすること、他の評価指標についての調査をすること、ウェブページの評価指標に関する既存研究の調査を行うことなどが挙げられる。

「参考文献」

- 1) S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine”, In Proc. 7th International World Wide Web Conference, (1998) pp.107-11.
 - 2) J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM, 46, 5, (1999) pp.604-632.
 - 3) D. Easley and J. Kleinberg, 『ネットワーク・大衆・マーケット -現代社会の複雑な連結性についての推論-』(浅野孝夫・浅野泰仁), 共立出版, (2013).
- *1) <https://dumps.wikimedia.org/jawiki>
*2) <https://dumps.wikimedia.org/other/pagecounts-ez>