企業の Web サイトにおけるアクセスログデータを用いた閲覧者の重視するページの解析に関する研究

Research on Analysis of Pages that are Important to Viewers Using Access Log Data on Corporate Websites

小迫良輔 †, 森山真光 †
Ryosuke Kosako†, and Masamitsu Moriyama†
† 近畿大学大学院 総合理工学研究科
†Graduate School of Science and Engineering Research, Kindai Univ.

要旨

近年，企業の製品情報を発信する場として Web サイトによる情報発信は重要な位置を占めている．Web サイトの閲覧者がどのようなコンテンツを重視しているのかを, Web サイトの管理者や企業の担当者が容易に知ることが可能になることを目的として, アクセスログを解析し, 得られた閲覧者の動向を考慮した PageRank アルゴリズムを利用し, Web サイト内の重要なページを判別する手法の開発と実験を行った．閲覧者の遷移を利用した PageRank を用いることで, PageRank が高いページから多く遷移されたページがより PageRank が高くなるため, 対象の期間中にユーザーが閲覧する可能性の高いページを推測することができると考えた.

1. はじめに

近年，企業の製品情報を発信する場として Web サイトによる情報発信は重要な位置を占めている．2019 年の国内の企業のホームページ開設率は高く, 全体の 89.7% となっている [1]．また, 2020 年初旬から始まった新型コロナウイルスの影響により, 実店舗の休業やオンラインショッピングの需要の増加に伴い, Web サイトに合わせた経営戦略がより必要な状況となっている．そのため企業が Web サイトに取り組む事に関しては重要な位置を占めており, Web 閲覧者の動向に合わせたコンテンツの変更や改善などを行う事が必要になっている．その方針を定める方法の一つして, Web サイトのアクセスログを用いたアクセスログ解析が挙げられる．従来サービスとして, Google 社の Google Analytics などがある．また, Web ページの重要度を決定するためのスコアリングアルゴリズムとして有名なアルゴリズムとして, Google で使用されている PageRank アルゴリズムがある [2]．このアルゴリズムでは, Web サイトのリンク構造を利用しており, リンク元のページの重要度と, 対象の Web ページへの被リンク数に基づいて Web ページごとの重要度を計算している．しかし, PageRank アルゴリズムでは閲覧者のアクセス履歴とは関係なく, 全てのリンクを等価に扱っているため, 時期や閲覧者の動向を考慮した重要度を計算することはできない.

そこで本研究では, ある企業の Web サイトの閲覧者がどのようなコンテンツを重視しているのかを, Web サイトの管理者や企業の担当者が容易に知ることが可能になることを目的として, Web ビーコン型のアクセスログ取得スクリプトから得られたアクセスログを解析し, 得られた閲覧者の動向を考慮した PageRank アルゴリズムを利用し, 企業の Web サイト内の重要なページを判別する手法の開発と実験を行った.

2.PageRank の概要

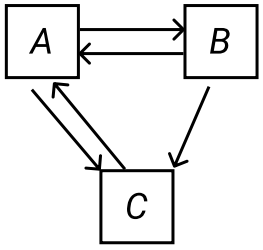
PageRank は Google の共同創設者である Larry Page と Sergey Brin によって 1998 年に提案されたアルゴリズム [2] で, リンク構造を元に Web ページの品質を評価する．PageRank では, より多くのページからリンクされたページのページランクが高くなり, また, リンク元のページのページランクもリンク先のページランクに影響する.

2.1.PageRank の計算方法

P を Web ページ全体の集合とする．$N(P_j)$ でページ P_j のリンク先の総数を示す．$B(P_i)$ で Web ページ P_i のリンク下の Web ページの集合を表す．α は現在のノード内のリンクを辿って遷移する確率とする．α の値は PageRank の考案者によって 0.85 が推奨されている．$R(P_i)$ をページ P_i の PageRank 値としたとき, 式 (1) となる．式 (1) を反復法を用いて繰り返し計算することで PageRank のスコアを求めることができる.

$$R(P_i) = \alpha \sum_{P_j \in B(P_i)} \frac{R(P_j)}{N(P_j)} + \frac{(1-\alpha)}{N} \tag{1}$$

実際に計算するときは, 式 (1) を行列計算に置き換える. 例えば, 図 1 のようなページ A, B, C のハイパーリンク構造の場合を考えたとき, 図 1 のハイパーリンクの構造を PageRank の行列計算に適した形にするため, 各ページごとの総リンク数で割った行列 S は式 (2) となる.



$$S = \begin{pmatrix} 0 & 1/2 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix} \tag{2}$$

図 1: ハイパーリンクの構造の例

行列 S を用いた計算は以下の式 (3) で表すことができる.

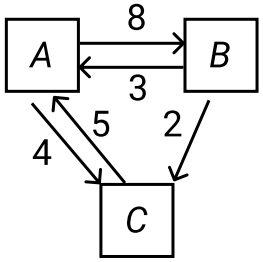
$$R(P_i) = \alpha \sum_{j=1}^{n} S^{\top} R(P_j) + \frac{(1-\alpha)}{N} \tag{3}$$

式 (3) をべき乗法を用いて R が収束するまで繰り返し計算を行うことで, 各ページの PageRank を求めることができる.

2.2. アクセスログを用いた PageRank の計算方法

従来の PageRank 値は各ページに対して均等に割り当てられる. 本研究ではユーザーのアクセスしたリンクの割合に応じて各ページへの PageRank 値を変える.

図 2 のようなページ A, B, C のアクセスのリンクの関係が成り立っている場合を考える. 矢印の値は訪問回数を示す. 例えば図 2 中の $A \to B$ の場合, ページ A からページ B へ 8 回の訪問があったことを示す. 図 2 のアクセスログのリンク構造を示す行列 S は式 (4) となる.



$$S = \begin{pmatrix} 0 & 3/5 & 1 \\ 2/3 & 0 & 0 \\ 1/3 & 2/5 & 0 \end{pmatrix} \tag{4}$$

図 2: アクセスログのリンク構造の例

式 (4) で示した行列 S を式 (3) に適用することで, 閲覧数を用いた PageRank 値を導出する.

3. 本研究で利用するデータ

本研究では, ある企業 A の協力で, 同社が運営している Web サイトに Web ビーコン型のページログ取得スクリプトを設置し, 収集されたデータを用いてアクセスログを加工して本研究の実験に用いる.

ページログ取得スクリプトによって, Web 閲覧者がアクセスした時刻 (UnixTime)，Web 閲覧者の IP アドレス，アクセスされたページの URL, 流入元 URL, CookieID, ブラウザのユーザーエージェント, ホスト等を取得し, アクセスログの管理サーバーに保存する.

また, Web ページのリンク構造の実データはクローラを用いて収集したデータを元に作成した.

4. 実験結果と考察

今回の調査では, 期間ごとの PageRank 値を計算し, 期間によって傾向が見られると仮定して実験を行った. また, 期間ごとに各ページの閲覧数の抽出と順位付けを行い, アクセスログを用いた PageRank と比較した. クローリングによって取得した 10732 ページとアクセスログを用いて PageRank 値を計算した. 表 1 に実験に利用したデータの期間と閲覧数, ユニーク URL 数を示す.

期間	閲覧数	ユニーク URL 数
2019 年 3 月 1 日から 2019 年 3 月 14 日	84682	2959
2019 年 7 月 1 日から 2019 年 7 月 14 日	76540	2999

表 1: 実験に利用したデータの期間, 閲覧数, ユニーク URL 数

表 2 は 2019 年 3 月 1 日から 2019 年 3 月 14 日の期間を対象とした実験結果である. 3 月は就職活動の時期のため, 採用情報関連のページの閲覧数が高くなった.

PageRank 順位	閲覧数順位	PageRank	閲覧数	Web ページ
1	2	0.038547971	3201	Web サイトトップページ (SSL 対応なし)
2	3	0.0245845633	2476	採用情報トップページ
3	4	0.019640053	2095	製品情報 (トップページ)
4	1	0.0192809545	7405	Web サイトトップページ
6	8	0.0170849289	1252	会社概要
10	16	0.0084113186	649	採用情報 (子ページ)
15	6	0.0072373427	1712	会社概要 (SSL 対応なし)
76	17	0.0022375864	633	事業 A(トップページ)

表 2: 2019 年 3 月 1 日から 2019 年 3 月 14 日までの実験結果

表 2 上の会社概要ページは SSL 対応なしのページは閲覧数では 6 位だが,PageRank は 15 位となった. 対して, SSL 対応された会社概要ページは閲覧数では 8 位だが PageRank は 6 位となっている. 企業 A のページは SSL 対応されていないページから SSL 対応されたページへのリダイレクトを行っていない. しかし, 多くのページは SSL 対応されたページへのリンクを設置しているため, SSL 対応されていないページは被リンク数が少なくなるため, 閲覧数が多くても SSL 対応されていない場合は PageRank が低くなる傾向が見られた.

表 3 に 2019 年 7 月 1 日から 2019 年 7 月 14 日の期間を対象とした実験結果を示す.

PageRank 順位	閲覧数順位	PageRank	閲覧数	Web ページ
1	3	0.0340107798	2590	Web サイトトップページ (SSL 対応なし)
2	2	0.0224363789	2691	会社概要
3	6	0.021946875	1426	会社情報 (トップページ)
4	1	0.0212520951	5871	Web サイトトップページ
9	13	0.0104830072	753	Web サイトトップページ (中国語)
30	7	0.0054932246	1088	事業 B(子ページ)
44	61	0.0033457232	214	事業 B(トップページ)
47	67	0.0031404473	201	事業 C(中国語)
60	79	0.0025404102	164	事業所一覧 (中国語)

表 3: 2019 年 7 月 1 日から 2019 年 7 月 14 日までの実験結果

表 2, 表 3 ともに, SSL 対応していない Web サイトトップページの方が PageRank が高くなった. 企業 A の Web サイトは検索エンジンでは多くのページは SSL 対応していないページがインデックスされている. 本実験では同サイト内での遷移から解析するため, 検索エンジンからの流入は PageRank の計算から除外されている. ヘッダーに設置されたトップページへのリンクは絶対パスではなく, 相対パスの”/”を指定している. トップページ以外から流入した閲覧者は検索エンジンから流入したページを閲覧したあと, ヘッダーのリンクからトップページに遷移する場合が多かった. そのため, 計算する際に SSL 対応されていないトップページへの流入割合は多くなるために PageRank が高くなったと考えられる.

表 2 の事業 A（トップページ）や表 3 の事業 B（子ページ）は閲覧数順位に対して PageRank 順位が著しく低い. このような場合, 検索エンジンからの流入は多く見込めるが, サイト内の巡回によって対象のページに辿り着きにくいことが考えられる. 検索エンジンからの流入が多いが, 実際には Web サイト内の巡回によっては閲覧される回数は少ないため, ユーザーは巡回中に対象のページ関係のコンテンツを目的に巡回はしていないと考えられる. そのため, より高い PageRank のページに関係するページ群はユーザーが Web サイト内を巡回中に求めているコンテンツを含んでいると考えられる.

企業 A は日本語の他に中国語のページを公開している. 表 3 の Web サイトトップページ (中国語), 事業 C(中国語), 事業所一覧 (中国語) のように中国語のページは総じて閲覧数順位よりも PageRank 順位が高くなる傾向が見られた. 2019 年 7 月 1 日から 2019 年 7 月 14 日の期間中の日本語と中国語のページ数と閲覧数を比較した結果を表 4 に示す.

言語	ページ数	閲覧数	1 ページあたりの閲覧数
日本語	2690	71130	26.4424
中国語	96	2982	31.0625

表 4: 言語別のページ数と閲覧数

表 4 に示すように, 1 ページあたりの閲覧数が日本語のページと比較して中国語のページの方が高くなるため, 中国語のページではより少ないページ間を訪問者が巡回していることになる. そのため, 中国語のページが総じて PageRank が閲覧数よりも高い傾向になったと考えられる. また, 本稿で示した手法では, 外部からの流入による閲覧数は PageRank に影響せず, Web サイト内を巡回されたことによる閲覧数が多いほど PageRank が高くなる. そのため, 巡回されたページの総数がより少ないほど全体的に PageRank が高くなるため, 中国語圏からの訪問者は日本語のページを閲覧する訪問者よりも同サイト内を巡回する回数の平均値が高いことが考えられる.

5. 結論

本稿では, ある企業の Web サイトの閲覧者がどのようなコンテンツを重視しているのかを, Web サイトの管理者や企業の担当者が容易に知ることが可能になることを目的に, 閲覧者の動向を考慮した PageRank アルゴリズムを利用し, 企業の Web サイト内の重要なページを判別する手法の開発と実験を行った. 実験結果の考察において, 閲覧者の遷移を利用した PageRank を用いることで, PageRank が高いページから多く遷移されたページがより PageRank が高くなるため, 対象の期間中にユーザーが閲覧する可能性の高いページを推測することができると考えた. 今後の課題として, SSL 対応していないページやルートフォルダを index.html としてリンクされたページなどを計算しやすいように URL の正規化をした場合の手法の提案や, カテゴリーを基準とした PageRank の改良が挙げられる.

参考文献

[1] 総務省, ”通信利用動向調査　令和元年通信利用動向調査　企業編　2019 年”, 政府統計の総合窓口, https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200356&tstat=000001140586&cycle=0&year=20190&month=0&tclass1=000001140587, (参照 2020-11-05)

[2] L. Page and S. Brin and R. Motwani and T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web, 1998