2P-6

時系列リンク解析を用いたWebページの評価指標に関する研究*

金子 圭一朗　古瀬 一隆　陳 漢雄†
筑波大学 システム情報系‡

1　はじめに

検索エンジンのランキングでは, PageRank[1] に代表される Web ページを評価する様々な指標が用いられる. こういった指標では一般的に, Web をある単一の時点におけるスナップショットとして切り取り, その中での Web ページ間のリンク構造を解析することで各 Web ページを評価している. これは HITS[2] や SALSA[3] といった評価指標でも同様である. そのため, これらの評価指標では, ページやリンクがいつ作成されたかなどといった時間的な要素については考慮していない. 例えば, クエリを「ワールドカップ　開会式」として検索すると, これを書いている現時点で, 前回大会のワールドカップに関連した Web ページばかりが上位に表示されてしまう. この場合, ユーザは次回大会に関連した Web ページを見たがっていると考えるのが一般的である. しかし, こういった Web ページについては, 現在の Web 構造しか見ていない従来の評価指標では価値を判断しにくい.

そこで本研究では, 過去から現在までの Web の時系列データを用いてリンク解析を行うことにより, Web グラフの動的な性質を考慮した Web ページの評価指標を定義する.

2　時系列リンク解析

Web の動的な性質を考慮してリンク解析を行うために, ある期間における Web データを逐一保存していき時系列順に並べた蓄積データを用いる（Web アーカイブと呼ばれる）. この Web アーカイブを用いて一定期間毎の Web のスナップショットを作成し, 複数のスナップショットを用いて Web の変化を考慮しながらリンク解析を行うことを, 本研究では時系列リンク解析と呼ぶ. これにより Web ページやリンクの作成・削除日時を取得して Web の移り変わりを捉えることで, より有効な Web ページの評価が行うことができると考えられる.

*A Study on Metrics of Web Pages using Time Series Link Analysis
†Keiichiro Kaneko, Kazutaka Furuse, Hanxiong Chen
‡Faculty of Enginnering , Information and Systems, University of tsukuba

3　提案手法

本研究では, 現在における Web ページの評価指標として, 被リンク数の増減に基づく評価指標 F と, 過去の評価値の履歴に基づく評価指標 R を提案する.

3.1　被リンク数の増減に基づく評価指標 F

最近注目を集めるようになってきている Web ページは現在において価値がある Web ページだと考え, 高く評価する. 例えば, 現在における評価値が同じ Web ページでも, 過去から現在にかけて評価値を上げている Web ページの方が, 過去から現在にかけて評価値を下げている Web ページより情報が新しく, 現在において価値がある Web ページだと判断する. 以上を踏まえて, 被リンク数の増減に基づく評価指標 F は以下の式 1 で表される.

$$F(t,p) = \int_{t' \le t} w(t') \Big(\sum_{q \in mk(t',p)} \frac{F(t',q)}{outdeg(t',q)} - \sum_{q \in rm(t',p)} \frac{F(t',q)}{outdeg(t',q)} \Big) dt' \quad (1)$$

ここで, $w(t)$ は t が小さいほど値が小さくなる減衰係数を表す. また $mk(t',p)$ は時刻 t' に新たに Web ページ p にリンクを張った Web ページ集合, $rm(t',p)$ は時刻 t' に Web ページ p へのリンクを削除した Web ページ集合を表す. また $outdeg(t,q)$ は Web ページ q の時刻 t における出リンク数を表す.

3.2　過去の評価値の履歴に基づく評価指標 R

過去から現在にかけて支持続けているような Web ページから多くリンクされている Web ページには価値があると考え, 高く評価する. 例えば, 現在における評価値が同じ Web ページでも, 長年安定して高く評価されている Web ページからリンクされているページの方が, 何らかの理由で一時的に評価を上げた Web ページからリンクされているページより信頼性が高く, 現在において価値ある Web

ページだと判断する. 以上を踏まえて, 過去の評価値の履歴に基づく評価指標 R は以下の式 2 で表される.

$$R(t,p) = \sum_{q \in i(t,p)} \int_{t' \leq t} \frac{w(t')R(t',q)}{outdeg(t',q)} dt' \qquad (2)$$

ここで, $w(t)$ は t が小さいほど値が小さくなる減衰係数を表す. また $i(t,p)$ は時刻 t において Web ページ p にリンクを張っている Web ページ集合を表す. また $outdeg(t,q)$ は Web ページ q の時刻 t における出リンク数を表す.

なお, 実際の計算では t' に下限を設けて収束値を求める. また, 評価値は Web アーカイブ上での Web ページの記録時刻による離散値の和での近似値となるため, Web アーカイブの性能によってはかなりの誤差が起きると考えられる. そのため, 実装上の計算では, Web アーカイブでの記録時刻の集合を $T = \{t_0, t_1, \cdots, t_n\}$ とするとき, 以下の式 3 で近似計算を行う.

$$R(t,p) = \sum_{i=1}^{n} \Big(\frac{w(t_i)R(t_i,q)}{outdeg(t_i,q)} + \frac{w(t_{i-1})R(t_{i-1},q)}{outdeg(t_{i-1},q)} \Big) \frac{t_i - t_{i-1}}{2} \qquad (3)$$

4 実験と評価

予備実験として過去の評価値の履歴に基づく指標 R についてのみ評価実験を行った. 19 種類の検索クエリで Google で検索を行ったランキング上位約 500 件を評価対象の Web ページとし, 各ページの過去の記録を代表的な Web アーカイブである InternetArchive[4] を用いて 1 年ごとに合計 10 年分 (2002-2011) を取得して, これらの蓄積データに対して時系列リンク解析を行い指標 R によって各 Web ページを評価した. また比較実験として, 評価対象の Web ページに対して PageRank[1] を適用させ, それぞれの指標で Web ページのランキングを作成した. 各手法におけるランキング上位 10 件に対して, 現在におけるそのページの価値という本研究の判断基準に基づき, 3 段階 (価値がある:2 点, どちらともいえない:1 点, 価値が無い:0 点) の採点をつけた.

実験結果を図 1 に示す. なお横軸は各クエリ, 縦軸は 20 点を最高とする評価である. また, 減衰係数については, n 年前の Web ページの場合 $(\frac{1}{2})^n$ となるようにした. 図 1 より, 指標 R を用いた場合の方が安定して良い結果となっていることが分かる. しかし, 扱った実験データが少ないため, 有効性を示すにはまだ不十分であると考えられる. より正確な有効性を図るためにさらなる実験が不可欠である.

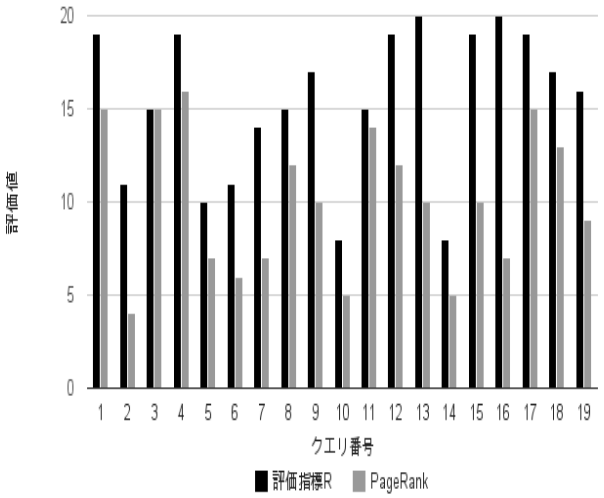

図 1: 指標 R についての予備実験

5 まとめ

本研究では Web の動的な性質を考慮し, 時系列リンク解析を用いて Web ページを評価する 2 つの指標を提案した. 予備実験の結果, 指標 R について良好な結果が得られた. 今後の課題として, 指標 F についての実験や, F と R を組み合わせた場合についても検討する必要がある. また実験データの規模・評価の方法についても検討中である.

参考文献

[1] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Standord Digital Library Technologies Project, 1988.

[2] J. M. Kleinberg. Authorative Sources in a Hyperlinked Environment. In Proc. of hte 9th Annual ACM-SIAM Symposium on Discrete Algorithm, pp.668-677, 1999.

[3] R. Lempel and S. Moran. SALSA: The Stochastic Approach for Link-structure Analysis. ACM Transactions on Information Systems, pp.131-160, 2001.

[4] Internet Archive. The Internet Archive. 2011 http://www.archive.org/.