様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　１　日現在

機関番号：　１２１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2012～2015
課題番号：　２４５００１０５

研究課題名（和文）時系列リンク解析に基づく重要度尺度に関する研究

研究課題名（英文）Research on Inportant Degree based on Temporal Link Analysis

研究代表者
古瀬　一隆（Furuse, Kazutaka）

筑波大学・システム情報系・准教授

研究者番号：１０２９１２８８

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：本研究では、Webページの評価指標として、以下の2つの評価指標を考案した。評価指標Fは被リンク数の増減に基づくものであり、評価指標Rは過去の評価値の履歴に基づくものであり、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ページだとの考えに基づき、そのようなページを高く評価する指標となっている。これに対し、評価指標Rは、過去から現在にかけて支持され続けているような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の考えに基づき、そのようなページを高く評価するものとなっている。これらの指標の有効性は、実データを用いた実験で確認した。

研究成果の概要（英文）：This research proposes two important degrees on web pages based on temporal link analysis. The degree F is measured with the increasing and decreasing of past values of important degrees. The degree R is measured with increasing and decreasing of in-links. The effectiveness of these degrees are confirmed by experiments with real data sets.

研究分野：　データ工学

キーワード：　リンク解析

１．研究開始当初の背景

リンク解析は Web グラフの構造を解析することによって様々な知識を抽出する手法であり、内容を記述している言語に依存せず、また、リッチコンテンツのようなテキスト情報を持たない Web ページにも適用できることなどから、さまざまな分野で研究されている。その中で代表的なものとして、Web ページの重要度尺度の決定や Web の関連ページの発見のための手法が研究されてきた。重要度尺度は Web ページの重要性を特定の基準に基づき数値化したものであり、よく知られたものには PageRank がある。

PageRank は Web ページの内容にたよらずリンク構造のみによって重要度を決定する尺度であり、Web ページの作者によって意図的に操作しづらいことから、サーチエンジンの検索結果の順位付けなどに用いられている。しかし、この手法では特定の時点の(通常は最新の)Web グラフを解析するに留まっており、このため状況によっては利用者が必要としている情報が高く評価されないことがある。例えば、研究開始時点において、ACM SIGMOD 2008 のページは数百の入リンクを受けているが、同 2011 のページは数十にすぎなかった。このような新しく作られたページは入リンク数を得るチャンスが少なく、PageRank などの既存の手法では必要以上に低く評価されることが有りうる。本研究では時系列リンク解析によってこのような問題の解決を図ることとした。

なお、PageRank とは異なり、問い合わせ内容に依存した重要度尺度として知られている HITS や SALSA などは、PageRank よりも精度がよいとの報告もあるが、これらの手法は問い合わせ時にリンク解析処理を行う必要があるため、実用上十分な応答速度を実現するのが困難であるという問題がある。この問題は、時系列リンク解析ではより顕著な問題となると考えられた。

２．研究の目的

本研究では、時系列 Web グラフに対してリンク解析することにより、既存の手法にはない新たな知識獲得手法を確立することを目的とした。現時点の(最新の)Web グラフを解析するだけでは得られないような知識としては、例えば以下のようなものがある。

- 現時点での評価が同じである Web ページであっても、過去により高い評価を得ていたにも関わらずその評価が減少して現時点の評価になったページと、最近になって評価を上げて現時点の評価になったページでは、後者のページの方が価値が高い。
- 現時点での評価が同じである Web ページであっても、長年にわたって高い評価を得続けているページと一時的に高い評価を得たページとでは、前者の方が価値が高い。

前項で例として挙げた ACM SIGMOD 2011 のようなページは、入リンク数は少ないが、短い期間に入リンク数を増加させているページである。また、SIGMOD 全体のトップページのような、長年にわたって高い評価を得続けているページからのリンクを受けている。したがって、このような価値を正当に評価する重要度尺度を確立することは新たな知識獲得手法として意味があると考えられる。このような知識を獲得するためには、過去から現在に至る時系列の Web グラフに対するリンク解析が必要となる。本研究ではこのような新たな価値に基づく Web ページの重要度尺度の確立を目指した。また、Web クラスタリングの効率化に必要となる近傍検索の基礎技術の検討もあわせて行った。

３．研究の方法

本研究ではまず、時系列リンク解析に基づく Web の重要度尺度についての検討を行った。リンク解析についてはこれまでにもさまざまな手法が提案されているが、それらは主として現時点での Web グラフのみを解析の対象としている。本研究では過去の Web グラフの履歴を用いて、各 Web ページの入リンク数等の増減の経緯に基づく新たなモデルと重要度尺度を定義した。

以上の検討により構築した手法・機構の有効性を評価するため、プロトタイプシステムを実装した。過去の Web グラフの履歴の取得には、世界最大級の Web アーカイブである Internet Archive(www.archive.org)のデータおよび独自に開発したクローラを用いて収集したデータを用いた。さらに、Web ページ間の近傍検索を行うことで重要な役割を果たす Web ページを効率的に発見するための基礎技術についても検討を重ねた。

また、HITS や SALSA のような手法では、問い合わせが与えられた際に計算に用いる Web グラフを決定し、それから処理を行うため、PageRank などの問い合わせ非依存型の手法に比べて応答速度が遅くなるという問題がある。そこで本研究では、あらかじめ前処理の段階で Web グラフを一定の規模以下の小さな部分グラフに分割し、これに対してハブ値やオーソリティ値を計算しておくことで、問い合わせ処理時の計算量を減らし、応答速度を向上させる手法を考案した。部分グラフに分割した場合のハブ値やオーソリティ値を用いて Web グラフ全体のハブ値やオーソリティ値を見積もる場合には誤差が生じることになるため、応答速度と精度はトレードオフの関係になることが予想されるが、Web コミュニティの抽出手法などを応用することにより、精度を保ったまま問い合わせ時の処理速度を向上させる手法を検討した。

４．研究成果

本研究の手法は、Web の動的な性質を考慮してリンク解析を行うものである。そのため、一定の期間における Web データを定期的に取得して時系列順に並べた蓄積データを用いている（このような蓄積データのことを本研究では Web アーカイブと呼ぶ）。この Web アーカイブを用いて一定の期間毎の Web のスナップショットを作成し、その複数のスナップショットを用いて Web のリンク構造の変化を考慮しながらリンク解析を行っている。これにより、Web ページやリンクの作成・削除の時刻を考慮した評価指標を実現する。

本研究では、Web ページの評価指標として、被リンク数の増減に基づく以下の 2 つの評価指標を考案した。評価指標 F は被リンク数の増減に基づくものであり、評価指標 R は過去の評価値の履歴に基づくものである。

評価指標 F では、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 ページだとの考えに基づき、そのようなページを高く評価する。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挙げている 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 ページであると考えられる。この考え方に基づいて考案した評価指標 F は以下の式で表される。

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

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,p)は時刻 t における Web ページ p の出リンク数を表す。

評価指標 R では、過去から現在にかけて支持され続けているような 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の考えに基づき、そのようなページを高く評価する。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 ページよりも信頼性が高く、現在において価値あるページだと判断する。この考え方に基づいて考案した評価指標 R は以下の式で表される。

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

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。outdeg(t,p)は時刻 t における Web ページ p の出リンク数を表す。

これらの指標について、本研究で新たに実装した Web クローラによって収集した実データを用いた被験者実験を行い、その有効性を確認した。図 1 はその結果の一部である。

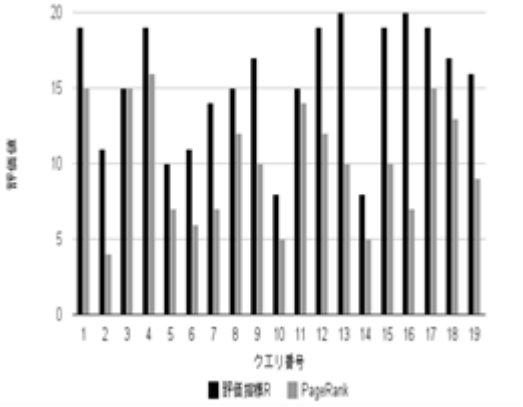

図 1 実データによる被験者実験の結果

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 4 件）

① 董 于洋, 陳 漢雄, 古瀬 一隆, 北川 博之. A Branch-and-Bound Method for Group Reverse Queries, 第 8 回データ工学と情報マネジメントに関するフォーラム, 2016-3-2, ヒルトン福岡シーホーク（福岡県福岡市）.

② Atsuhiro Ichikawa, Hanxiong Chen, Kazutaka Furuse. Efficient reverse far neighbors search, Ninth International Conference on Digital Information Management, 2014-9-30, Phitsanulok (Thailand).

③ 市川 敦啓, 陳 漢雄, 古瀬 一隆. 逆遠方検索とその効率的な検索方法に関する研究, 第 6 回データ工学と情報マネジメントに関するフォーラム, 2014-3-3, ウェスティン淡路（兵庫県淡路市）.

④ 金子圭一朗, 古瀬一隆, 陳漢雄. 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, 情報処理学会第 74 回全国大会, 2012-3-6, 名古屋工業大学（愛知県名古屋市）.

６．研究組織

(1)研究代表者

古瀬　一隆（Furuse, Kazutaka）
筑波大学・システム情報系・准教授
研究者番号：10291288

(2)研究分担者

陳　漢雄（Chen, Hanxiong）
筑波大学・システム情報系・講師
研究者番号：60251047