インクリメンタル PageRank による重要 Web ページの効率的な収集戦略

山田 雅信† 田浦 健次朗† 近山 隆‡
東京大学大学院情報理工学系研究科電子情報学専攻† 東京大学大学院新領域創成科学研究科基盤情報学専攻‡

1 はじめに

近年のインターネットの爆発的な普及に伴い、Web 上には多種多様かつ膨大な情報が存在するようになった。しかし、そのような情報の中から何の手がかりも無しに欲する情報のみを見つけ出すことは不可能に近い。そのため現在では、情報検索の効率化のためにサーチエンジンと呼ばれる情報検索システムが広く利用されている。

このようなサーチエンジンは、その性格上、大きく２つに分けることができる。それは、Web リソースの選別と収集、そしてインデックス構築といった作業の大部分を人手によるディレクトリ型と、これら一連の作業をプログラムによって自動的に行なうロボット型である。

特に後者において，Web リソースを自動収集するプログラムはクローラ、(WWW)ロボット、スパイダーなどと呼ばれる。そして、クローラによって収集される Web ページ数は膨大であるため、検索時のユーザからの検索要求に対する検索結果も膨大なものとなってしまうことが多々ある。そのため、検索結果を Web ページの重要度に基づき的確にランク付けし、それにより検索結果をソート、提示することはサーチエンジンにとって非常に重要な要素となる。

しかし、従来のクローラの多くは幅優先探索などのように Web ページの収集順序にこのような重要度を考慮していない。また、例え考慮していたとしても、それらは単純な局所情報によるところが大きかった。そのため、検索時のランク付けにおいて低いランクしか持たないような利用価値の低い Web ページまで収集してしまい無駄が多いと言える。一方、Web ページ数が指数関数的に増加しているのに対し、さまざまな理由によりクローラの収集能力がそれに追随できなくなってきていることなどから、未収集の Web ページの中に利用価値の高い重要なものが収集されずに残っているという可能性も無視できない。

そのため、重要な Web ページを効率的に収集するようなクローリング戦略が重要となるが、これに対し Web ページの収集順序の決定に、検索時のランク付けアルゴリズムとして用いられる PageRank の利用[1]などが考えられる。しかし、その計算方式はインクリメンタルな収集データに対しては不向きであり、より低コストの計算方式が必要となる。そこで本研究において、新たな Web ページ収集戦略としてインクリメンタル PageRank を提案、実装するとともに、クローリングのシミュレーション実験を行い、その有効性を示すことを目的としている。

本稿ではまず、２章においてインクリメンタル PageRank の基本となる従来の PageRank の概念、３章において本研究における収集戦略ついて述べ、最後に４章でまとめを行う。

Efficient Collection Strategies of Important Web Pages by Incremental PageRank

† Department of Information and Communication Engineering, Graduate School of Information Science and Technology, University of Tokyo
‡ Department of Frontier Informatics, Graduate School of Frontier Sciences, University of Tokyo

2 PageRank

PageRank[2][3]は、ページＡからページＢへのリンクをページＡによるページＢへの支持投票とみなし、この投票数により投票された側のページの重要度を評価する。しかし、単に投票数のみで評価するのではなく、票を投じたページに対しても評価を行い、その一票一票に対し重み付けを行う。つまり、重要度の高いページからの票は、より重要なものとして扱う。あるページの重要度は、この重みの総和により決定される。

この関係を具体的に見ると図 1 のようになる。投票する側のページの PageRank を、そのページのもつ他ページへのリンクの総数で割った値が、それぞれリンク先のページへ forward され、投票される側では forward された値の総和がそのページの新たな PageRank となる。しかし、この考えをそのまま実際の Web グラフに適用すると、そのリンク構造によって PageRank 計算に支障をきたすことが知られている。そこで実際の PageRank 計算においては、各ページに対し次式を一度だけ適用するといった作業を作業単位としたとき、その作業単位を繰り返し行うことにより真の PageRank である収束値を求めている。

$$PR(a) = d + (1-d)\sum_{i=1}^{n} PR(p_i)/N(p_i)$$

ここで、PR(x)はページ x の PageRank、p_i はページ a へのリンクをもつページ、$N(p_i)$はページ p_i のもつ他ページへのリンクの総数、d は damping factor と呼ばれるパラメータで 0.15 とい

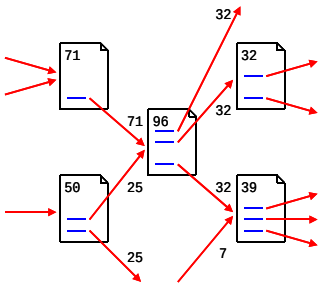

図１ PageRank の概念

う値がよく用いられる。

3　本研究における Web ページ収集戦略

3.1 収集アルゴリズムとインクリメンタル PageRank

クローラによる Web ページ収集の基本動作は

1. クローリングのスタート地点として与えられた URL に該当するページを収集する
2. 新たに収集したページに含まれているリンクを列挙し、まだ収集していないページへのリンクがあったら“未収集 URL リスト”に登録する
3. “未収集 URL リスト”から次に収集すべきページを決定し、そのページを収集する
4. 条件を満たした時点で終了となる。そうでなければ２に戻る

のようになる。ここで重要となるのが“次に収集すべきページ”を“どのように”決定するかである。後々の検索のためにも利用価値の高い、すなわち重要なページを効率的に収集することが望ましい。本研究ではそのようなページを PageRank の高いページとし，その効率的収集を考える。

上記３が終了した時点で、収集済みのページに関してはそのリンク構造が既知であるため、その PageRank を求めることができる。また、“未収集 URL リスト”のページはこの時点ではまだ dangling page として扱われているが、どの収集済みページからリンクされているかは既知であるため、先ほど求めた PageRank とリンク情報をもとに、“未収集 URL リスト”のページに対しても PageRank の予想値を求めることができる。これにより“未収集 URL リスト”のページの収集優先順位を決定し、最も優先順位が高いものを実際に収集する。収集されたページはさらにその次に収集するページの決定に利用される。

ここで問題となるのが PageRank の計算方式である。新たにページを収集したことにより、今まで自分が保持していたリンク構造が局所的に変化するが、ここで厳密な全ページの PageRank を求め直すことはコストが大きい。そこで本方式では PageRank の更新範囲をリンク構造が変化した部分を中心とした一部分とし、収集と更新を繰り返すことにより PageRank の収束を図る。この方式をインクリメンタル PageRank と呼ぶこととする。

3.2 インクリメンタル PageRank の更新方法

収集アルゴリズムの２において、PageRank の更新がどのように行われるかは、新たに収集したページがリンクしているページの分類によって大きく異なる。もし新たに収集したページが未知のページや、“未収集 URL リスト”内のページへリンクしていた場合はそれほど深刻な問題とはならない。新たに収集したページから PageRank の計算式どおりの PageRank をリンク先に forward するだけでよく、それによる収集順序の入れ替えも起こり得るが、それは小規模なものである。

一方、新たに収集したページが既に収集したページへリンクしていた場合、高精度の PageRank を望むなら、新たに収集したページからリンクを辿って到達可能なページ全てに対し PageRank の更新を行う必要がある。これは非常にコスト高である。しかし、実際は、あるページが多くのリンクを持っているとき、新たに PageRank を forward することによるリンク先での PageRank の更新の影響はそれ以降急激に小さくなる。このことからも、新たにページを収集したことにより forward される PageRank による PageRank の更新を途中で打ち切ることで、その更新にともなう計算の大幅な削減が望める。現段階においては以下のような場合に打ち切るようにしている。

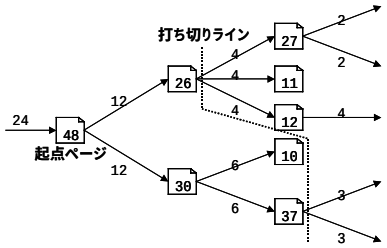

図２ 新たに forward する PageRank による打ち切り例（閾値５）

- 新たに forward する PageRank がある閾値以下（図２）
- forward される側のページの PageRank がある閾値以下
- 新たに forward する PageRank と forward される側のページの PageRank 比がある閾値以下
- 更新されたページ数がある閾値以上
- 起点からのリンクの深さがある閾値以上

その詳細な打ち切りについては今後の実験により改善していく。

4　おわりに

クローラが収集段階でページの重要度を判定し、重要なページを効率的に収集することは、結果として検索サービスの質の向上につながる。本稿では、そのためのクローラの収集戦略を扱った。しかし、インクリメンタル PageRank においては当然ながら真の PageRank との間に誤差が生じる。今後の主な課題として、さらなるインクリメンタル PageRank の更新範囲決定手法を提案するとともに、それらの手法による PageRank 計算の高速化率、そのとき失われる計算精度のトレードオフ関係の測定と改善が挙げられる。

参考文献

[1] J.Cho, H.Garcia-Molina, L.Page, *Efficient Crawling Through URL Ordering*,, Proceedings of the 7th International World Wide Web Conference, 1998.

[2] L.Page, S.Brin, R. Motwani, T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford Digital Librairies Working Paper, 1998.

[3] S.Brin and, L.Page, *The Anatomy of a Large-scale Hypertextual Web Search Engine*, Proceedings of the 7th International World Wide Web Conference, 1998.