ISS-P-008

PageRank のダンピングファクターの値による計算量の傾向の分析

小迫　良輔†　　森山　真光†
† 近畿大学大学院 総合理工学研究科

1. はじめに

PageRank [1]は検索エンジンに使われている中で最も有名なスコアリングアルゴリズムの一つである．このアルゴリズムは Web ページが持つリンク構造を基に各ページの重要度をスコアリングするもので，基本的な考え方は重要なページは重要なページにリンクされているという考え方から成り立っている．PageRank には利用範囲に応じて適切なパラメータを追加することで，ノード間の関連度の計算に用いることが可能であり，グラフマイニングにおいて有用な手法であるが，計算量が多くなるという特徴がある．その計算をする上で，ダンピングファクターという値を設定することで，計算の反復回数の調整が可能である．そこで，最適なダンピングファクターを求めるためにダンピングファクターの値の違いによる計算量の傾向について分析する．

2. PageRank とダンピングファクター

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．基本概念はネットサーフィンをするユーザーをモデル化したランダムサーファーモデルに基づいている．ネットサーフィンを行うユーザーは無作為に各 Web ページにあるリンクをランダムに辿ると仮定し，各 Web ページに辿り着く確率からスコアを決定する．ダンピングファクターは，ページ内のリンクを辿って別のページに遷移する確率である．ダンピングファクターは 1 に近づくほど反復して計算する回数は増えるが，その値はPageRank の考案者によって 0.85 という値が計算を行う上で現実的に妥当な数字であったと述べられている[2]．そのため，多くのPageRank の計算ではその値が用いられている．

3. 実験結果

図 1 はそれぞれノード数を 100, 1000 の場合でダンピングファクター(α)の値を 0.05 から 0.95 まで 0.05 ずつ増やし，0.05 ごとの計算の反復回数の増加率を計測した結果である．α が 0.65 未満の場合は増加率が安定しなかった．0.65 以上からはどちらのノード数の場合も増加率は同じ傾向となった．また，α が 0.8 以上の場合，増加率は著しく増加した．α が 0.8 以上では著しく計算量が増加するため，推奨される α = 0.85 を採用していた場合，計算量を減らすために α を 0.8 未満にすることが一つの有効な手段として考えられる．

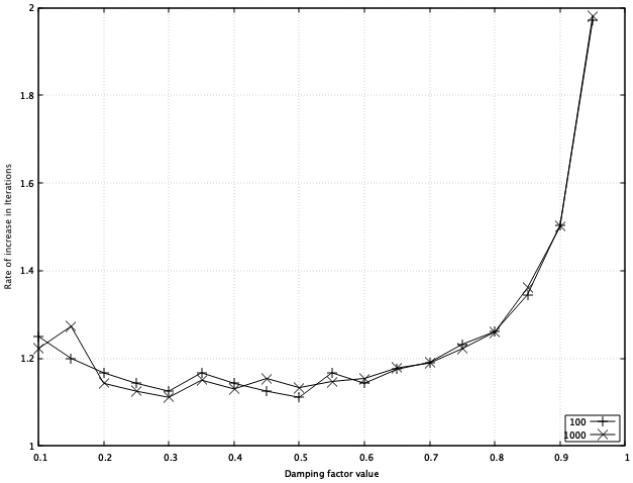

図1. ダンピングファクターの値による反復回数の増加率

また，図2は α を 0.8, 0.85, 0.9, 0.95, 0.99 のそれぞれノード数を100 ずつ増やした計算時間の結果である．α が大きいほどノード数が増えたときに計算時間の増加が大きいことが分かった．

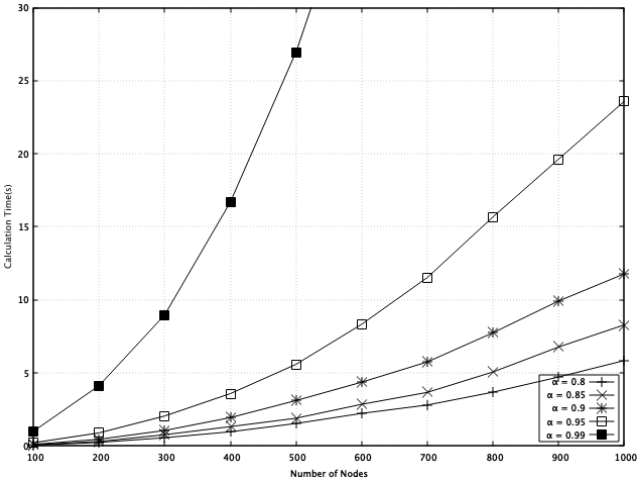

図 2. ノード数ごとの計算時間(α=0.8, 0.85, 0.9, 0.95, 0.99)

4. まとめ

今後はテストデータを用いて PageRank の計算時間を事前に計測し，その結果から最適なダンピングファクターを予測する手法を考察する予定である．

参考文献

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

[2] Amy N. Langville and Carl D. Meyer. Google's Pagerank and Beyond: The Science of Search Engine Rankings. Princeton Univ Pr, 7 2006.