Webページ間最短経路探索システムの構築

松永　拓[†]　平手　勇宇[††, †††]　山名　早人[††††, †††††]

[†]早稲田大学大学院基幹理工学研究科　[††]早稲田大学大学院理工学研究科

[†††]早稲田大学メディアネットワークセンター　[††††]早稲田大学理工学術院　[†††††]国立情報学研究所

1 はじめに

直接リンクされないWebページ間のつながりを知ることは、Web構造を把握するために重要である。1999年のBarabasiらの研究[1]によれば任意の2ページ間の平均距離は19であると推定されている。しかし現在、リンク情報を収集しているWeb検索エンジンにおいても任意のページ間のリンクによるつながりを知る機能はなく、ページを探索して経路を発見するのも困難である。

そこで、本稿では、任意のWebページ間のつながりを高速に得ることを目的に、我々が2006年9月に収集した約1億ページの日本語Webページを用いて、Webページ間最短経路探索システムを構築する。また、本システムを構築のために、本データを対象として最短経路アルゴリズムの評価を行う。

2 最短経路アルゴリズム

本システムで用いる最短経路アルゴリズムについて述べる。事前計算の必要のない両方向探索と、事前計算の必要なALTアルゴリズム[2]を用いて比較を行う。

2.1 両方向探索

ダイクストラ法[3]をスタート、ゴールの両端からお互いが同じノードを訪れるまで探索する手法である。

2.2 ALTアルゴリズム

Goldbergらによって提案されたALTアルゴリズム[2]は、A*探索[4]を基としたアルゴリズムである。A*探索においては、2点間の距離の下限値を推定することにより、問題を最適化する。しかし、地理情報などの距離推定のための付加情報がないグラフにおいては、どのように下限値を推定するかが問題である。ALTアルゴリズムでは、下限値の推定に、事前に選択し全点への距離を計測したノードへの距離情報を用いる。この事前に選択されたノードをランドマークと呼ぶ。

2.3 ALTアルゴリズム + 両方向探索

池田らは双対変数を用いることでA*探索を、両方向探索と組み合わせる手法[5]を提案している。[2]では、この手法を応用し、ALTアルゴリズムを両方向探索に拡張している。

A System for Finding Shortest Paths Between Web Pages

Taku Matsunaga[†], Yu Hirate[††, †††], Hayato Yamana[††††, †††††]

[†]Graduate School of Fundamental Science and Engineering, Waseda University

[††]Graduate School of Science and Engineering, Waseda University

[†††]Media Network Center, Waseda University

[††††]Science and Engineering, Waseda University

[†††††]National Institute of Informatics

3 ALTアルゴリズムでのランドマーク選択

ALTアルゴリズムにおいて、良いランドマークを選択することが重要である。今回は、[2]で提案している選択法のうち、次の2手法を用いる。

3.1 Random選択

ランドマークをランダムに選択する。事前の計算コストがもっとも少ないのが利点である。

3.2 Farthest選択

最初にランダムに頂点を選び，その頂点からもっとも遠い頂点を選択し，ランドマークに加える。以降、同様にランドマーク集合から一番遠い頂点を選び、ランドマークに加える。ただし、Webでは他の多くのページへの経路がないページがあるので、そのようなページは候補から外した。

5 最短経路アルゴリズムの実験

5.1 実験データ

本実験には我々が2006年9月に収集した日本語Webページから、正リンク、逆リンクの両方を1つ以上持つページのみを抽出して用いる。このデータには、113678448のページと3571838753のリンクが含まれる。

本データを、[6]で提案されたグラフフォーマットで変換したファイルの情報を表1で示す。正方向リンク、逆方向リンクの2セットが必要である。また、ランドマークファイルは、1つのランドマークにつきノードファイルと同じファイルサイズとなる。

表1. グラフファイルのサイズ

ノードファイル	**433.65MB**
エッジファイル	**13625.48MB**

5.1 実験条件

実験では，(1) 両方向探、(2) ALTアルゴリズム(Farthest)、(3)ALTアルゴリズム(Random)+両方向探索、(4) ALTアルゴリズム(Farthest) +両方向探索、の4つの手法についてそれぞれアルゴリズム中のノードの訪問回数と実行時間を比較する。ランダムに選択したスタートノード、ゴールノードを10セット作成し、各アルゴリズムの実験には同一のセットを用いて計測を行い、平均の訪問回数、実行時間を実験結果とする。実行時間の測定には、ディスクキャッシュの影響を考慮し、10セットの試行を複数回実行し、最終の値を採用する。

ALTアルゴリズムでは事前にランドマークとの距離を計測する必要がある。正方向リンク、逆方向リンクにおいて、それぞれに16個ずつ選択し、計算を行った。

実行マシンは、CPUはXeon 2.66GHzの8Core、メモリは16GBである。ただし、並列処理は行っていない。

5.2 実験結果

表 2 に実験結果を示す。

表 2. 探索終了までのノードの訪問回数

	訪問回数	実行時間(秒)
(1)両方向探索	24223.57	5.57
(2)ALT(Farthest)	3148539	5029.16
(3)ALT(Random)+両	14031.2	3.06
(4)ALT(Farthest)+両	14120.3	3.06

6 実際の Web ページ間最短経路

本節では、構築した本システムを用いて、探索を行うと具体的にどのような最短経路が得られるのかを示す。表 3 は、日本語 Google トップページから、Yahoo! Japan への最短経路である。「Google と Yahoo! JAPAN が提携 - Google の検索エンジン採用で機能を拡充」という 2001 年 4 月 2 日付けの Google のプレスリリースを中継して、Yahoo! Japan にたどりつけることがわかる。なお、補足ではあるが、2004 年 5 月 31 日に Yahoo! Japan は Google の検索エンジンの利用を終了している。

表 4 は、読売新聞社のトップページから朝日新聞社のトップページ、表 5 は、朝日新聞社のトップページから読売新聞社のトップページへの最短経路である。どちらも他サイトを通じてつながっており、距離は 4 ホップであった。

図 1 は、本システム上で、経路の存在するランダムな 2 点を 50 セット選択し、距離を計測した結果である。平均距離は 13.64 であり、表 3、表 4、表 5 で示したサンプルが、十分に短い距離であり、関連性があるということがわかる。

表 3. Google.co.jp から Yahoo.co.jp への最短経路

距離	タイトル / URL
0	Google トップページ *http://www.google.co.jp/*
1	Google 人材募集 *http://www.google.co.jp/intl/ja/jobs*
2	プレスセンター *http://www.google.co.jp/intl/ja/press/*
3	プレス リリース *http://www.google.co.jp/intl/ja/press/pressreleases.html*
4	Google と Yahoo! JAPAN が提携 *http://www.google.co.jp/intl/ja/press/pressrel/pressrelease4.html*
5	Yahoo Japan!トップページ *http://www.yahoo.co.jp/*

表 4. Yomiuri.co.jp から Asahi.com への最短経路

距離	タイトル / URL
0	YOMIURI ONLINE *http://www.yomiuri.co.jp/*
1	熊本 : 地域 : YOMIURI ONLINE *http://www.yomiuri.co.jp/e-japan/kumamoto/*
2	リンク : 熊本 : 地域 : YOMIURI ONLINE *http://www.yomiuri.co.jp/e-japan/kumamoto/link/*
3	熊本県西原村公式ホームページ *http://www.vill.nishihara.kumamoto.jp/*
4	朝日新聞の速報ニュースサイト *http://www.asahi.com/*

表 5. Asahi.com から Yomiuri.co.jp への最短経路

距離	タイトル / URL
0	朝日新聞の速報ニュースサイト *http://www.asahi.com/*
1	朝日新聞社会社案内/グループ企業・関連団体 *http://www.asahi.com/shimbun/honsya/j/associate.html*
2	朝日オリコミ *http://www.asaori.co.jp/*
3	朝日オリコミ - 折込お役立ち情報 >新聞折込広告関連リンク集 *http://www.asaori.co.jp/orikomi/link.html*
4	YOMIURI ONLINE *http://www.yomiuri.co.jp/*

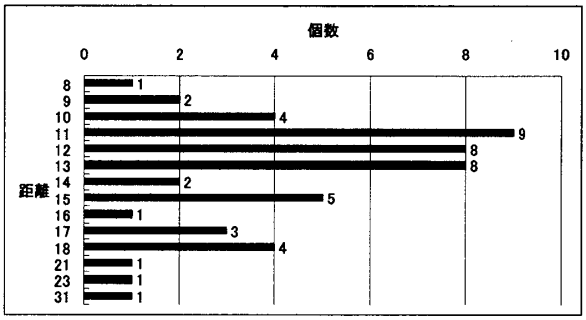

図 1. ランダムに選択した 2 点間の距離

7 まとめ

本稿では、両方向探索と ALT アルゴリズムを、約 1 億ページの Web データに対して適用し、両方向探索 ALT アルゴリズムでは実用的な時間で探索が可能なことを確認した。また、構築したシステムを用いて、3 例の探索を示し、Web ページ最短経路により、2 ページ間の関連性を把握できることを確認した。

参考文献

[1] R. Albert, H. Jeong and A.-L. Barabasi, "Diameter of the World-Wide Web," Nature, Vo. 401, No. 9, pp. 130-131, 1999.

[2] A. V. Goldberg and C. Harrelson, "Computing the Shortest Path: A* Search Meets Graph Theory," In Proc. 16th ACM-SIAM Symposium on Discrete Algorithms, pp. 156–165, 2005.

[3] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," Numer. Math., 1:269–271, 1959.

[4] P. E. Hart, N. J. Nilsson, and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," IEEE Transactions on System Science and Cybernetics, SSC-4(2), 1968.

[5] T. Ikeda, Min-Yao Hsu, H. Imai, S. Nishimura, H. Shimoura, T. Hashimoto, K. Tenmoku, and K. Mitoh, "A Fast Algorithm for Finding Better Routes by AI Search Techniques," In Proc. Vehicle Navigation and Information Systems Conference. IEEE, 1994.

[6] A. V. Goldberg and R. F. Werneck, "Computing Point-to-Point Shortest Paths from External Memory," In Proc. 7th International Workshop on Algorithm Engineering and Experiments, pp. 26–40. SIAM, 2005.