数理実習B (離散手法) レポート

山西研究室サイトの利便性・保守コストの改善方法の考察

学生証番号: XX-XXXXXX

氏名: 佐藤 瞭

連絡先: XXX

概要

ウェブサイトを、必要な情報を得やすく保守もしやすい構造にしたいという問題がある。そこで、ウェブサイトををウェブグラフの形で表現し、最短路長、強連結成分、ページランクを求めることで、各特徴量からサイトの利便性や保守のしやすさを議論し、改善点を考えることを試みた。解析対象のデータは山西研究室のサイトとした。

背景･目的

ウェブサイトを、必要な情報を得やすく、保守がしやすいものにしたい、という問題がある。必要な情報を得にくいウェブサイトは使いづらいとみなされ利用者が減り、保守がしにくいウェブサイトは保守に高いコストがかかってしまうためである。そのため、ウェブサイトを必要な情報を得やすく、保守がしやすいものにする手法は、利用者の増加や保守コストの削減という観点において有用である。そこで、今回の実験の目標を、実際に存在するウェブサイトの離散的な構造を、ウェブページをノードに、リンクをエッジに見立てたウェブグラフの形で把握し、ウェブグラフの構造の特徴から、ウェブサイトの必要な情報の得やすさや保守のしやすさを議論し、改善点を考えることとした。

データ

数理情報工学コースの山西研究室 (http://ibis.t.u-tokyo.ac.jp/yamanishiken/) のデータを用いた。

データは、講義で与えられたクローラーのスクリプト (http://www.misojiro.t.u-tokyo.ac.jp/~denzumi/lectures/B3Experiment/web/crawler.py) を用いて2019/06/14に入手した。クローリングを実行する際は、http://ibis.t.u-tokyo.ac.jp/yamanishiken/ から到達可能な全てのURLを探索し、ドメイン外のURLに達したらOut of domainと記録してそれ以上深く探索しないようにした。

解析

- ウェブグラフの全ノード数は1005、全エッジ数は8575であった。
- ウェブグラフのファイルの種類ごとのノード数を表1に示す。「Out of domain」はhttp://ibis.t.u-tokyo.ac.jp/yamanishiken/ 以下にファイルが置かれていないことを、「error」はリンク切れなどでファイルを表示できなかったことを、「application」は画像とhtml以外のファイルを意味する。

表1 ウェブグラフのファイルの種類ごとのノード数

File Type	count
Out of domain	427
html	303
application	121
image	102
error	52

- htmlのノードのみで構成したウェブグラフのノード数は303、エッジ数は5138であった。図1は、このウェブフラフを可視化したものである。

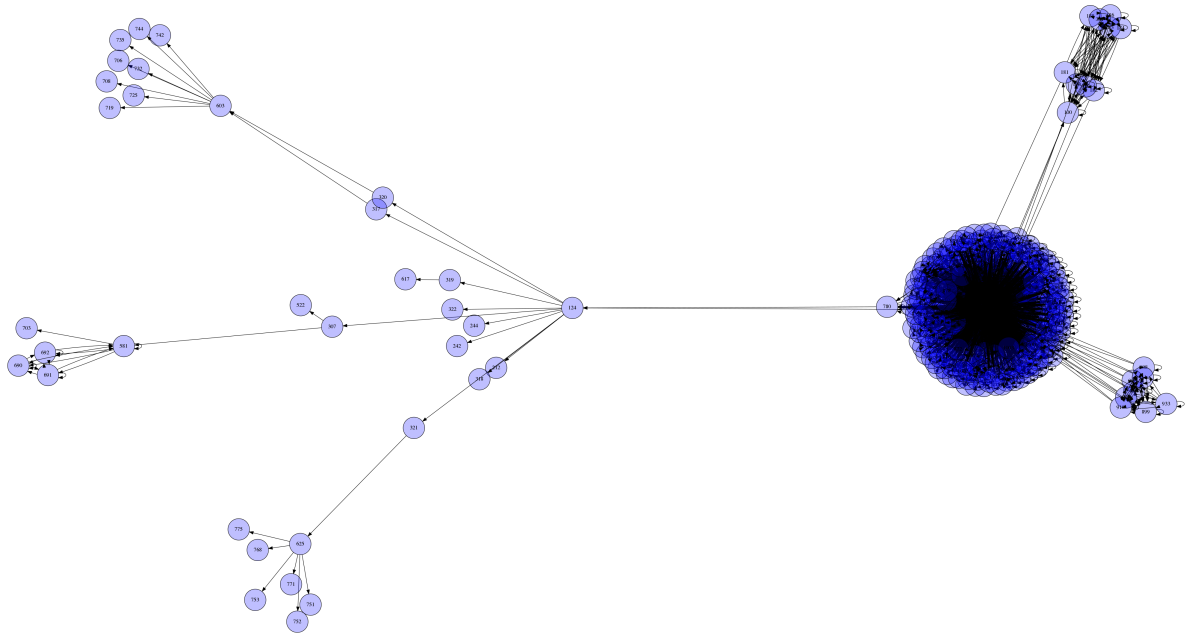

図1 htmlのノードのみで構成したしたウェブグラフ

以下では、図1のウェブグラフに対する解析をしていく。

- 図2に、ノード間の最短路長の分布を示す。最短路長の平均値は13.17であった。

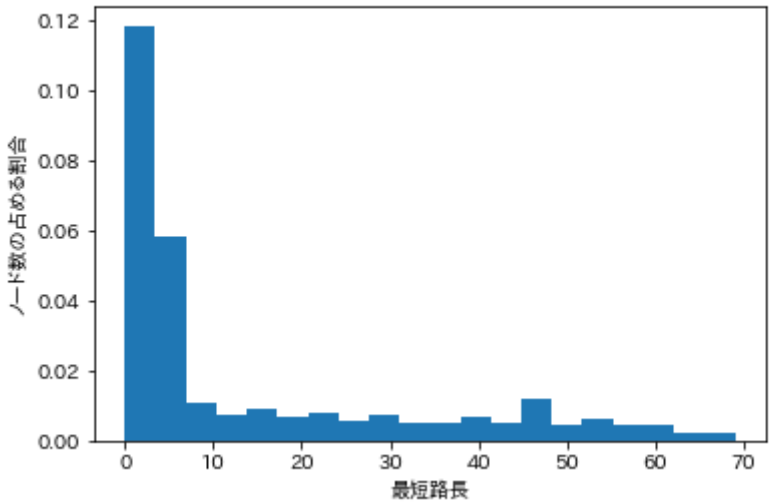

図2 ノード間の最短路長の分布 (平均値: 13.17)

- 強連結成分の数は32であった。また、ある1つの強連結成分は269個のノードを含み、別の1つの強連結成分は4個のノードを含み、他の全ての強連結成分は、それぞれがただ1つのノードを含んでいた。
- 表2に、Page Rankの最も大きい20ページを示す。図3に、Page Rankの値の分布を示す。

表2 Page Rankの最も大きい20ページ (Page Rankの値の大きい順に並べてある)

Title	Page Rank
山西研究室	0.045221
メンバー専用ページ | 山西研究室	0.042988
研究内容 | 山西研究室	0.042988
研究コンセプト | 山西研究室	0.042988
発表文献リスト | 山西研究室	0.042988
セミナー情報 | 山西研究室	0.042988
講義 | 山西研究室	0.042988
メンバー | 山西研究室	0.042988
CRESTについて | 山西研究室	0.042988
サイトマップ | 山西研究室	0.042988
KDD2019にTutorial: “Modern MDL meets Data...	0.001243
山西教授が日本脳ドック学会総会で特別講演を行います。| 山西研究室	0.001301
二反田助教がSummer School 2019 on Transfer Learningで...	0.001301
山西教授が編集を務めた数理科学6月号「特集:データサイエンスの数理」が刊行されました。| ...	0.001301
山西教授が電子情報通信学会サービスコンピューティング研究会で招待講演を行います。| ...	0.001301
KDD2019に論文が受理されました。| 山西研究室	0.001301
IJCAI2019に論文が受理されました。| 山西研究室	0.001301
山西教授が眼科AI・ビッグデータ研究会で招待講演を行います。| 山西研究室	0.001301
Data Mining and Knowledge Discovery誌に論文が採択されまし...	0.001301
研究室メンバーリストを更新しました。| 山西研究室	0.001301

考察

- 図1をみると、ノードをいくつかのクラスターに分類できるように見える。また、図1の左側には木に近い構造がみられる。この構造の根にあたるノードのURLはhttp://ibis.t.u-tokyo.ac.jp/suzuki/ (鈴木先生のホームページ)であった。このURLに直接アクセスできるのは、http://ibis.t.u-tokyo.ac.jp/yamanishiken/member/ およびhttp://ibis.t.u-tokyo.ac.jp/yamanishiken/en/member/ の2つ (研究室のメンバー紹介の日本語/英語版) のみであった。そのため、個人のサイトへのリンクは、研究室の「メンバー紹介」のページのみに記載されており、研究室のメンバーの入れ替わりがあった際に整理がしやすいように管理できているといえる。

- 図1の左側の木のような構造を持つサイトは、各節点での情報の分け方が明確で、かつその点よりも深いところにある情報をユーザーがわかるように表現してあれば、ユーザーが目的とする情報に高速に到達できるため、必要な情報を得やすいウェブサイトであるといえる。また、新規にページを追加する際はノードを1つ選び、それを節点として枝を1本伸ばすだけで木構造を維持できるため、保守コストは低いと考えられる。
- 最短路長の平均値は13.17であったが、図2をみると、最短路長の分布はロングテールになっている。最短路長が長いURLの組み合わせでは路の終点への到達が困難になってしまうため、サイトのリンクの張り方には改善の余地がある可能性がある。具体的な方法として、2つのノードの (順序付き) 組に重要度のスコアを定義して、組の重要度とその組の最短路長を引数とする目的関数を定義し、エッジの本数などの制約条件を与えたうえで、目的関数が最小化されるようにエッジの張り方を定める、といった処理が挙げられる。
- 表2を見ると、「山西研究室」「メンバー専用ページ」「研究内容」「研究コンセプト」「発表文献リスト」「セミナー情報」「講義」「メンバー」「CRESTについて」「サイトマップ」のページランクの値が、他のサイトと比べて極めて高い。そのため、これらのページ群は山西研究室のサイトにおいて最も重要であるといえる。これらのサイトはすべて、山西研究室のサイトのヘッダーメニューから直接到達可能である。そのために、これらのページの入次数が他のページと比べて極めて大きくなり、ページランクが大きくなることに繋がったと考えられる。これらのページに、他の重要なページのリンクを貼ることで、重要な情報への最短路を短くすることができ、必要な情報を得やすいウェブサイトにすることができるだろう。
- 強連結成分のうち4個のノードを含む成分は、ヘッダーつきの、セミナーの案内用サイト(http://ibis.t.u-tokyo.ac.jp/suzuki/ysg2014/index.html)であった。成分内の各ノード間を移動できるが、この成分の外に出るリンクは貼られていない。このページ群は1つのイベントに対する情報をもつものであり、ただ1つのノードを持つ強連結成分と同様の性質をもつといえるだろう。強連結成分のうち269個のノードを含む成分は山西研究室内のページであると考えられる。なぜならば、山西研究室内のページには共通のヘッダーやフッターが含まれており、ヘッダーやフッターのリンクから到達可能な山西研究室内のページは全て互いに到達可能となるためである。このような構造のサイトの場合、新規にページを追加する際はヘッダーやフッターのリンクから直接到達できるようにするか、ヘッダーやフッターのリンクから到達可能なページのいずれかにリンクを貼れば、新規ページを同じ強連結成分に含めることができ、新規ページはヘッダーやフッターから到達可能となる。しかし、これは新規ページへの到達しやすさにつながるわけではないため、リンクを貼る場所を適切に選ばないと新規ページへの最短路長が大きくなってしまう。したがって、このような構造は、先述の木構造と比べて保守コストは高いといえる。