Marisabel Chang Chan, Ocean Lu, Richard Pham, Derek Yee
CS4990 - Web Search
Dr. Ben Steichen
9 May 2019

Project 2 - Report

Introduction

The purpose of this project is to work with two important components of web search and search engines: Web Crawler and PageRank Calculation. The Web Crawler finds and downloads web pages by following links that are contained in the visited pages. The PageRank Calculation shows the rank of the most important web pages, people, citations, etc. in a collection.

Method (Code Analysis)

We separated our code into three different sections: The utility methods contains the functions that we use in the other two sections (Main driver and Print PageRank). In the main driver, we use the function to crawl and update the outlink table and PageRank list.

For the utility methods, we used different functions to parse the HTML text, update the outlink table, and check if the link is contained in the queue. To parse the HTML text, we use the get_html() function that returns the downloaded HTML texts only if the URL is a “text/html” type and the HTTP status is OK. The parse_html() function gets the links that are contained in the HTML document and returns a set that contains unique links. To update the outlink table, we created the updates_outlink_table method that adds a row and a column of zeros to the outlink table when a new page is visited and set 1 to the specific row (codomain) and column (domain) in the table. Then, it increments the number of outlinks to the page that the outlink comes from. Also, we used the divide_outlinks() function that divides all items of each column with its respective number of outlinks, so that the sum of all items of each column is equal to 1. For the checking functions, we used the has_link() function that checks if the queue contains the link already.

Finally, for the utility function, we have calculate_page_rank() that calculates the PageRank of each page contained in the graph. In this function, we give the same weight to each page in the PageRank list. Then, we do a matrix multiplication of the outlink table and the PageRank list until the previous PageRank list is equal to the actual PageRank.

In the Main Driver section, we request the HTML document, get the outlink list, and update the outlink table. Then, we check if each outlink is from the same base URL of the seed and if it has been crawled. If so, we increment one to the specific position in the outlink table. If not, we add the outlink to the queue or update the outlink list for the specific link. Finally, we divide each column a the specific number of outlinks.

In the Print PageRank section, we calculate the PageRank and display the top 100 pages that have higher PageRank.

Challenges

We encountered different challenges, such as retrieving the correct links to be parsed, updating the outlink table, and optimizing the runtime performance of the program.

In terms of getting the correct links, we had a problem that some links used alias names, causing outlink weights to be split. For example, cpp.edu and cpp.edu/index.shtml were considered two different pages, even though both links are linked to the same page. We fixed this problem by searching for the string “index” and removing it and everything after it. Also, we had a problem that we were concatenating a path of a URL to a URL that already contains the same path (ex. cpp.edu/~aboutcpp/~aboutcpp/). We fixed this problem by checking if the path was already contained in the URL before concatenating.

We encountered an issue when attempting to update the outlink table, where our table retrieved incorrect results from the web crawl. We solved this problem by creating a list that contained all the pages that land to the same page. Thus, every time a page is removed from the queue, the outlink table is updated with the changes.

Finally, we ran into another problem where some pages were taking more time checking if their outlinks were already visited or in the queue. Then, we realized that the list of the links that was returned in the parse_html() method was not unique. We solved this problem by changing the data structure of the list to a set. This optimized the amount of time it took to check outlinks, since duplicates do not exist in sets. Also, we changed the data structure of the visited_links variable to a set, too, for the same reason.

PageRanks

SEED: 'https://www.cpp.edu'

Number Pages Downloaded: 2500

Lambda: 0.20

Rank	Link	PageRank
1	https://www.cpp.edu	0.04525294055587124331
2	https://www.cpp.edu/~aboutcpp	0.04525294055587124331
3	https://www.cpp.edu/file-viewers.shtml	0.04521133778471275733
4	https://www.cpp.edu/~library	0.02367918239143323653
5	https://www.cpp.edu/accessibility.shtml	0.02236785926170458197
6	https://www.cpp.edu/website-feedback.php	0.02236785926170458197
7	https://www.cpp.edu/privacy.shtml	0.02236785926170458197
8	https://www.cpp.edu/events/day.php?date=04/19/2019	0.00792685254681482765
9	https://www.cpp.edu/~stratcomm/for-the-media.shtml	0.00373862069775227068
10	https://www.cpp.edu/~class	0.00346597183863431172
11	https://www.cpp.edu/~aboutcpp/why-cpp/well-ranked.shtml	0.00306330151733182586
12	https://www.cpp.edu/~aboutcpp/calpolypomona-overview/mission-and-values.shtml	0.00306277065477273888
13	https://www.cpp.edu/~aboutcpp/calpolypomona-overview/facts-and-figures.shtml	0.00305932128978270981
14	https://www.cpp.edu/~aboutcpp/why-cpp/affordable-and-accessible.shtml	0.00305932128978270981
15	https://www.cpp.edu/~aboutcpp/why-cpp/quality-learning-experiences.shtml	0.00305932128978270981
16	https://www.cpp.edu/~aboutcpp/calpolypomona-overview/public-good.shtml	0.00305932128978270981
17	https://www.cpp.edu/~aboutcpp/heritage/history-archives.shtml	0.00305932128978270981
18	https://www.cpp.edu/~aboutcpp/calpolypomona-overview/campus-community.shtml	0.00305932128978270981
19	https://www.cpp.edu/~aboutcpp/heritage/kellogg-house-pomona.shtml	0.00305932128978270981
20	https://www.cpp.edu/~aboutcpp/calpolypomona-overview/presidents-message.shtml	0.00305932128978270981
21	https://www.cpp.edu/~aboutcpp/heritage/kellogg-connection.shtml	0.00305932128978270981
22	https://www.cpp.edu/~agri	0.00275032687609386628
23	https://www.cpp.edu/~stratcomm/sharing-news.shtml	0.00266080444849505979
24	https://www.cpp.edu/~stratcomm/getting-news.shtml	0.00266080444849505979
25	https://www.cpp.edu/~stratcomm/contact-us.shtml	0.00262119628939905345
26	https://www.cpp.edu/~cba	0.00259770633231651196
27	https://www.cpp.edu/~class/news	0.00243586586860777311
28	https://www.cpp.edu/~ceis	0.00242422467303862223
29	https://www.cpp.edu/~alumni	0.00240611008942580827
30	https://www.cpp.edu/~class/departments.shtml	0.00234286470377767725
31	https://www.cpp.edu/~class/centers.shtml	0.00233857709872082199
32	https://www.cpp.edu/~class/contact-us	0.00233857709872082199
33	https://www.cpp.edu/~class/giving	0.00233857709872082199
34	https://www.cpp.edu/~class/alumni	0.00233857709872082199
35	https://www.cpp.edu/~class/about	0.00233857709872082199
36	https://www.cpp.edu/~admissions	0.00230084433245097411
37	https://www.cpp.edu/events/day.php?date=04/19/2019&audience=all	0.00228168339442923153
38	https://www.cpp.edu/~extended-education	0.00228150164806902684
39	https://www.cpp.edu/~giving	0.00214335725244028334
40	https://www.cpp.edu/~outreach/tours.shtml	0.00205053938013871834
41	https://www.cpp.edu/~aboutcpp/administration.shtml	0.00195504803411726893
42	https://www.cpp.edu/~sci	0.00194510332999734683
43	https://www.cpp.edu/~collins	0.00181985533679030398
44	https://www.cpp.edu/~alumni/outstanding-alumni/distinguished-alumni	0.00180132409527504890
45	https://www.cpp.edu/~commencement	0.00170074912942810982
46	https://www.cpp.edu/~agri/student-life	0.00163290246774064687
47	https://www.cpp.edu/~agri/giving	0.00163290246774064687
48	https://www.cpp.edu/~agri/about	0.00163290246774064687
49	https://www.cpp.edu/~agri/degrees	0.00163290246774064687
50	https://www.cpp.edu/~agri/contact-us	0.00163290246774064687

51	https://www.cpp.edu/~agri/alumni	0.0016329024677406468 7
52	https://www.cpp.edu/~agri/faculty-staff	0.0016329024677406468 7
53	https://www.cpp.edu/~agri/departments	0.0016329024677406468 7
54	https://www.cpp.edu/~officeofequity/titleIX	0.0016176511134653153 8
55	https://www.cpp.edu/~aboutcpp/visitor-information/dining.shtml	0.0015296154513212611 9
56	https://www.cpp.edu/~aboutcpp/visitor-information/parking.shtml	0.0015050960098139691 2
57	https://www.cpp.edu/~aboutcpp/visitor-information/attractions.shtml	0.0014797062479360962 34
58	https://www.cpp.edu/~aboutcpp/news.shtml	0.0014853059661002320 6
59	https://www.cpp.edu/~aboutcpp/visitor-information/shopping.shtml	0.0014853059661002320 6
60	https://www.cpp.edu/~aboutcpp/visitor-information/lodging.shtml	0.0014853059661002320 6
61	https://www.cpp.edu/~aboutcpp/visitor-information/directions.shtml	0.0014853059661002320 6
62	https://www.cpp.edu/~aboutcpp/why-cpp/diverse.shtml	0.0014853059661002320 6
63	https://www.cpp.edu/~aboutcpp/visitor-information/getting-around.shtml	0.0014853059661002320 6
64	https://www.cpp.edu/~cba/student-success-center	0.0014646892236135547 29
65	https://www.cpp.edu/~cba/academics	0.0014548877772074979 9
66	https://www.cpp.edu/~cba/college-giving	0.0014548877772074979 9
67	https://www.cpp.edu/~cba/alumni-friends	0.0014548877772074979 9
68	https://www.cpp.edu/~cba/about	0.0014548877772074979 9
69	https://www.cpp.edu/~cba/our-faculty	0.0014548877772074979 9
70	https://www.cpp.edu/~ceis/about	0.0013811680681077812 4
71	https://www.cpp.edu/~safety	0.0013733326076152852 4
72	https://www.cpp.edu/alpha-index	0.0013519307836245999
73	https://www.cpp.edu/campus-safety-plan.shtml	0.0013519307836245999
74	https://www.cpp.edu/annual-security-report.shtml	0.0013519307836245999
75	https://www.cpp.edu/~jobs	0.0013365562545926204 5
76	https://www.cpp.edu/contact.shtml	0.0013027103541935092 7
77	https://www.cpp.edu/~ceis/news-events	0.0012968299331251351 0
78	https://www.cpp.edu/~ceis/alumni-and-friends	0.0012968299331251351 0
79	https://www.cpp.edu/~ceis/contact-us	0.0012968299331251351 0
80	https://www.cpp.edu/~ceis/giving	0.0012968299331251351 0
81	https://www.cpp.edu/~ceis/departments.shtml	0.0012968299331251351 0
82	https://www.cpp.edu/~alumni/outstanding-alumni/spotlight/andrew-kopp.shtml	0.0012957161448900506 2
83	https://www.cpp.edu/events/index.php?cat=Cultural+and+Performing+Arts	0.0012857989140374118 4
84	https://www.cpp.edu/student-gateway	0.0012848335880612945 2
85	https://www.cpp.edu/~compass	0.0012830648188680202 6
86	https://www.cpp.edu/~alumni/about	0.0012829093605815193 3
87	https://www.cpp.edu/~alumni/contact	0.0012829093605815193 3
88	https://www.cpp.edu/~alumni/programs	0.0012829093605815193 3
89	https://www.cpp.edu/~alumni/student-outreach	0.0012829093605815193 3
90	https://www.cpp.edu/~alumni/chapters	0.0012829093605815193 3
91	https://www.cpp.edu/~aboutcpp/visitor-information?C=D;O=A	0.0012504249410269746 4
92	https://www.cpp.edu/~aboutcpp/visitor-information?C=M;O=A	0.0012504249410269746 4
93	https://www.cpp.edu/~aboutcpp/visitor-information?C=S;O=A	0.0012504249410269746 4
94	https://www.cpp.edu/~engineering	0.0012392168664550314 4
95	https://www.cpp.edu/~financial-aid	0.0012173266233024453 8
96	https://www.cpp.edu/~admissions/international?C=D;O=A	0.0011977280975595532 2
97	https://www.cpp.edu/~admissions/international?C=S;O=A	0.0011977280975595532 2
98	https://www.cpp.edu/~admissions/international?C=M;O=A	0.0011977280975595532 2
99	https://www.cpp.edu/~env	0.0011974244083532170 1
100	https://www.cpp.edu/faculty-staff-gateway	0.0011974244083532170 1

Contributions

Everybody contributed to writing the report, testing the calculations of PageRank by using a different number of pages and fixing bugs. Marisabel wrote the code to crawl and calculated the PageRank.