Final Report for Search Engine **Goodu**

Introduction

This is the base implementation of a full crawler that uses a spacetime cache server to receive requests. We have 4 people in our group. They are Bolun Sun, 52332355; Siheng Zhao 60161723; Tianxiong Wu 14266053; Jiaxiang Wang 39566477.

Choose of 20 Queries(5 for each):

Problem Queries:

1. David Redmiles
2. Pierre Baldi
3. UCI Alumni
4. Machine Learning
5. Michael Carey
6. Chen Li
7. Charless Fowlkes
8. Data Mining

Good Performance Queries:

1. The Transformative Play Lab
2. Noteworthy achievements
3. The Transformative Play Lab
4. Noteworthy achievements
5. video game development club
6. UCI DATASET
7. Computer Science
8. Dean’s List
9. Graudate School of UCI
10. job for the undergrad
11. UCI Spring 2020 Schedule
12. uci webreg

Chanllange and How to solve it?

2-GRAMS:
At first, the names of people always can't be searched effectively since many Chinese professors or people have Chen for their last name. Therefore, we use 2-grams to join the two words together and solve most of the names and two combinational words problem.

Page Rank:
Many trivial websites keep showing up when we are searching for Data Mining and Machine Learning. After a group meeting, we decided to use Page Rank to improve the quality of the website, which successfully returned the well-known and informative websites for students to learn Machine Learning and Data Mining.

Similarity(Simhash):
As the professor mentioned in the lecture, websites on the internet contain tons of duplicate information. Therefore we use simhash to deduplicate the website with a high similarity in order to provide the user 5 unique websites.

Sample Output

The Transformative Play Lab
1 . https://transformativeplay.ics.uci.edu/classes/
2 . https://transformativeplay.ics.uci.edu/arvr-theater-syllabus/
3. https://transformativeplay.ics.uci.edu/classes/inf-241-introduction-to-ubiquitous-computing/
4 . https://transformativeplay.ics.uci.edu/classes/ics-169-capstone/#schedule
5 . https://transformativeplay.ics.uci.edu/research/publications/

Noteworthy achievements
1 . https://www.ics.uci.edu/community/news/notes/notes_2014.php
2 . https://www.ics.uci.edu/community/news/notes/
3 . https://www.ics.uci.edu/~dan/class/267P/datasets/calgary/book1
4 . https://www.ics.uci.edu/~wscacchi/ResearchBio.html#OS

5 .
http://archive.ics.uci.edu/ml/support/Heart+Disease#cb68337ad074a5f5ce7d8ca8a7a0b7bad1931070

UCI DATASET

1 .
http://archive.ics.uci.edu/ml/support/Breast+Cancer#3e78257004181e6dbbdfa3ec12399520412e9c5c
2 .
http://archive.ics.uci.edu/ml/support/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks)#e5d994d772cfe5ec4d0f3e6d669f0bc28180a3ae
3 .
http://archive.ics.uci.edu/ml/support/Wine#a32ab1b3da96c9ae515a685b4fcf50e857708f24
4 .
http://archive.ics.uci.edu/ml/support/Heart+Disease#cb68337ad074a5f5ce7d8ca8a7a0b7bad1931070
5 .
http://archive.ics.uci.edu/ml/support/Abalone#ae82a44ada49c66439b67eae7ff10392ff209df9

video game development club

1 . https://transformativeplay.ics.uci.edu/arvr-theater-syllabus/
2 . https://www.ics.uci.edu/community/news/notes/notes_2014.php
3 . https://transformativeplay.ics.uci.edu/classes/ics-169-capstone/#schedule
4 . https://transformativeplay.ics.uci.edu/research/publications/
5 . https://www.ics.uci.edu/~eppstein/pubs/pubs.ff

Evaluation criteria:

- Does your search engine work as expected of search engines?
  - Closed but still a large distance compared to Google
- How general are the heuristics that you employed to improve the retrieval?

    - We use agile development and test the output while implementing the new features and filters
- Is the search response time under the expected limit?
    - Yes, the range of one search is 0.001 ~ 0.003 seconds
- Do you demonstrate in-depth knowledge of how your search engine works?
    - Yes, we used demo video to explain the code and how it works
- Are you able to answer detailed questions pertaining to any aspect of its implementation?
    - Yes.

Extra Credit:

1. **Detect and eliminate duplicate pages.** (1 point for exact, **2 points for near**)
2. **Add HITS and Pagerank to ranking.** (1.5 point HITS, **2.5 for PR**)
3. **Implement 2-gram and/or 3-gram indexing and use it during retrieval. (1 point)**
4. **Enhance the index with word positions and use that information for retrieval. (2 points)**
5. Index anchor words for the target pages (1 point).
6. **Implement a Web or GUI interface instead of using the console. (1 point for the local GUI, 2 points for a web GUI)**

Finished Extra Credit Question:

1. Detect and eliminate duplicate pages.	**+ 2**
2. Pagerank to ranking	**+ 2.5**
3. 2-gram indexing	**+ 1**
4. Enhance the index with word positions	**+ 1**
5. Web GUI	**+ 2**

Total Extra Credit: 8.5