Time-Travel on the Internet Via An Internet Archiving System

Group Members: Chan Lut Yan Loretta, Tang Siu Leung, Yeung Cheuk Yuen, and Yip Kai Ho Howard

Supervisor: Prof. Lin Gu

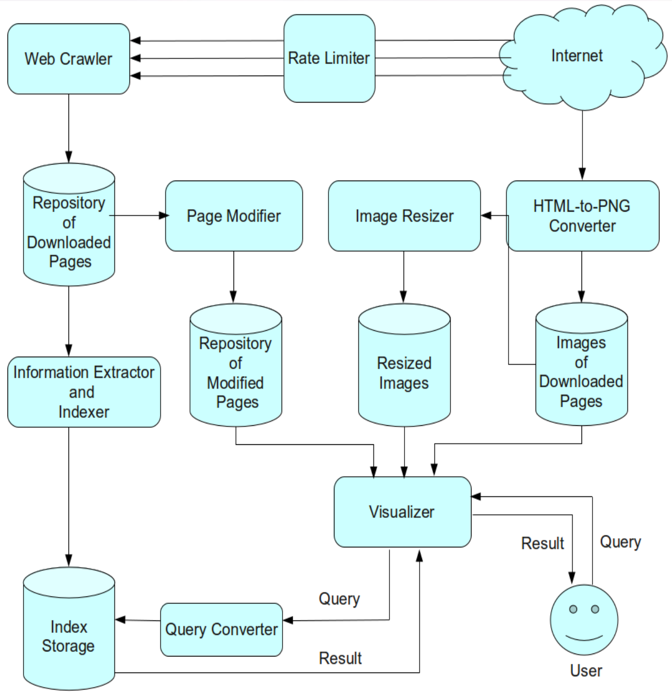
INTRODUCTION

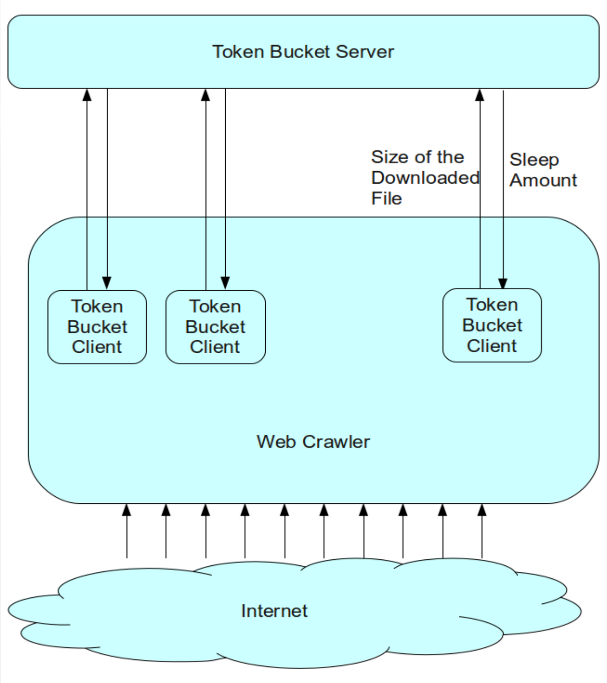
Internet browsing enables people to locate a myriad of information, but web pages are continually updated, so it is usually impossible to view archived web pages. This project implemented an Internet archiving system to allow virtual "time travel on the Internet."

DESIGN

The system includes several components:

1. Web crawler
2. Download rate limiter
3. Information extractor and indexer
4. Page Modifier
5. HTML-to-PNG converter
6. Query convert and visualizer



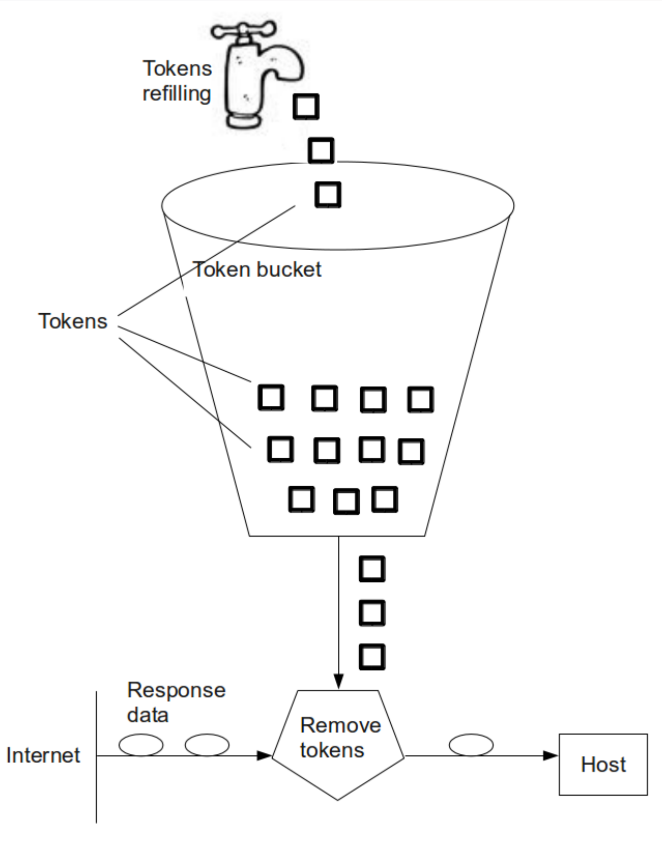


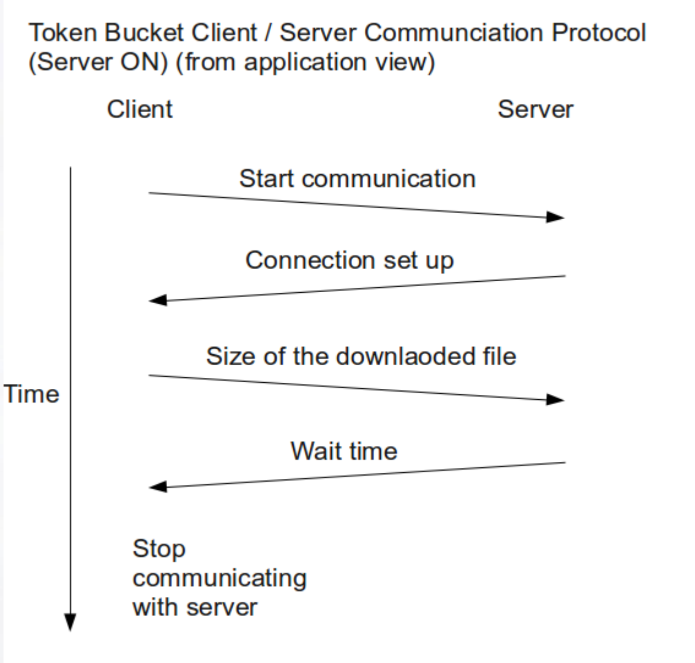
Web Crawler

- Automated parallel downloading
- Passive download rate limiting
- Scheduled seed site downloading
- Robot Exclusion Protocol Obedience

Rate Limiter

- Simulates network traffic shaping
- Implements the Token Bucket Algorithm
- Multi-threaded implementation of the algorithm
- Compatible with any program via transmission control protocol (TCP)




Token Bucket Client / Server Communciation Protocol (Server ON) (from application view)
Client
Server
Start communication
Connection set up
Size of the downlaoded file
Wait time
Time
Stop communicating with server

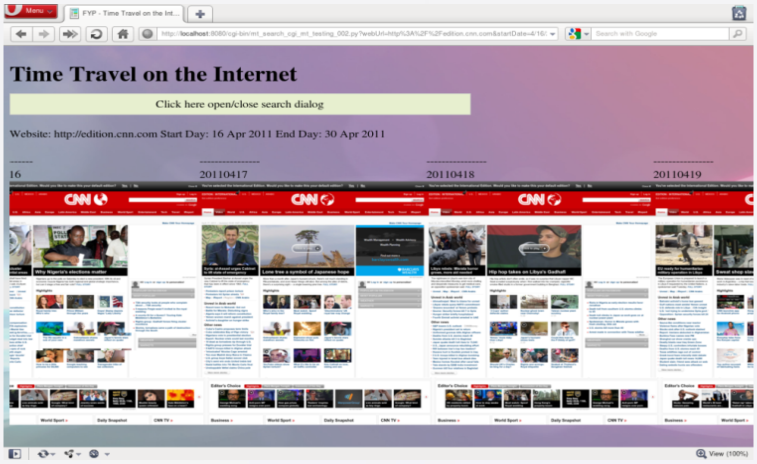
Information Extractor and Indexer

- Search by complete webpage addresses
- Search by part of webpage addresses
- Compare contents between 2 webpages

Page Modifier, HTML-to-PNG Converter

- Facilitating the visualizers with preserved appearances of different pages

Visualizers (Web Interface and Tk Interface)


Menu
FYP - Time Travel on the Int...
http://localhost:8080/cgi-bin/mt_search_cgi_mt_testing_002.py?webUrl=http%3A%2F%2Fedition.cnn.com&startDate=4/16/
Search with Google
View (100%)

Time Travel on the Internet

Click here open/close search dialog

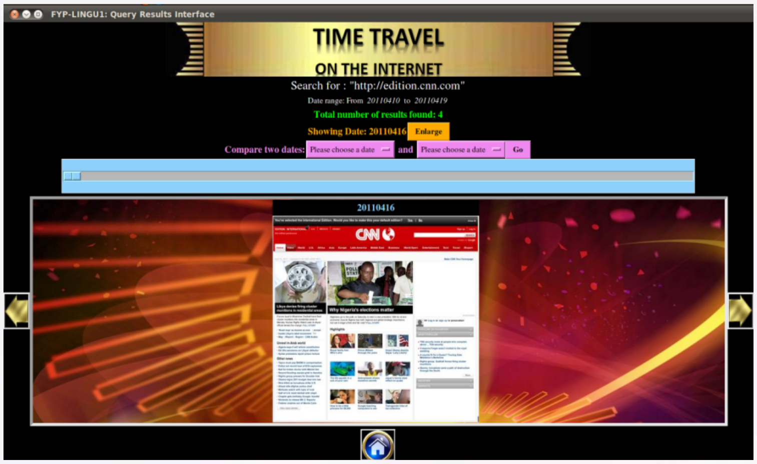
Website: http://edition.cnn.com Start Day: 16 Apr 2011 End Day: 30 Apr 2011

16

20110417

20110418

20110419


FYP-LINGU1: Query Results Interface

TIME TRAVEL
ON THE INTERNET

Search for : "http://edition.cnn.com"

Date range: From *20110410* to *20110419*

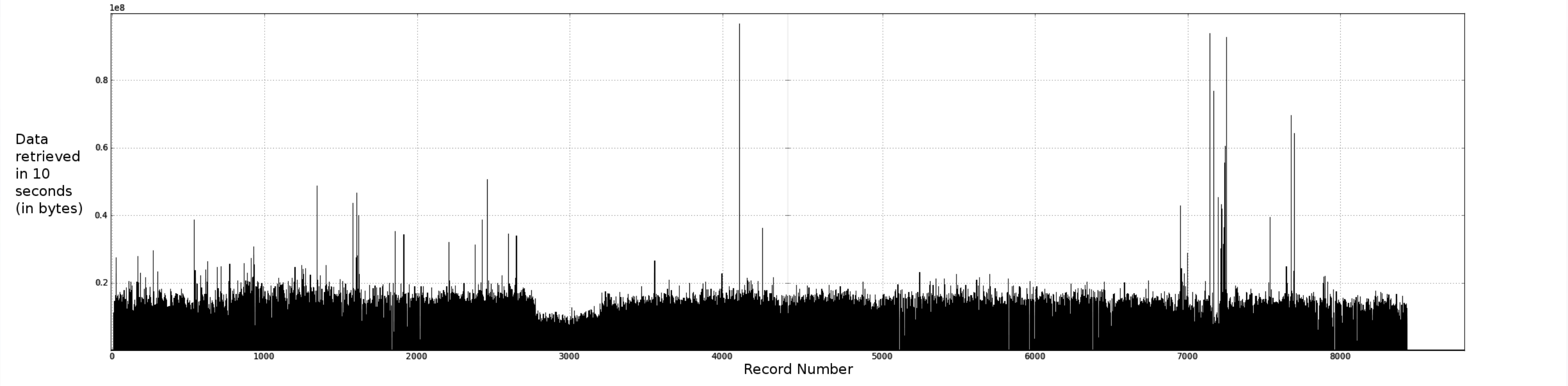
Total number of results found: 4

Showing Date: 20110416 Enlarge

Compare two dates: Please choose a date and Please choose a date Go

20110416

STATISTICS



Download rate captured every 10 seconds

CONCLUSION

In this project, we built a system that saves, converts, indexes pages, and display archived webpages. The system allows users compare webpages over time conveniently. The system also allows developing extensions easily.