World Repo - The repository of the World

Presented By

V. Srikrishnan
V. Shunmugachamy
CDAC - Chennai

What is a Search Engine ?

- A computer program that retrieves documents or files or data from a database or from a computer network.
- Uses a technique called 'Crawling'.
- Crawling is followed by,
- Indexing
- Inject->add more URL to Database
- Generate ->lists URL
- Parsing
- Invert links

World Repo – An Introduction

- Python based search Engine
- Gives active URL for download
- Existing Search Engines, gives all the links (active & dead links)
- Overhead for the User
- World Repo resolves User overhead

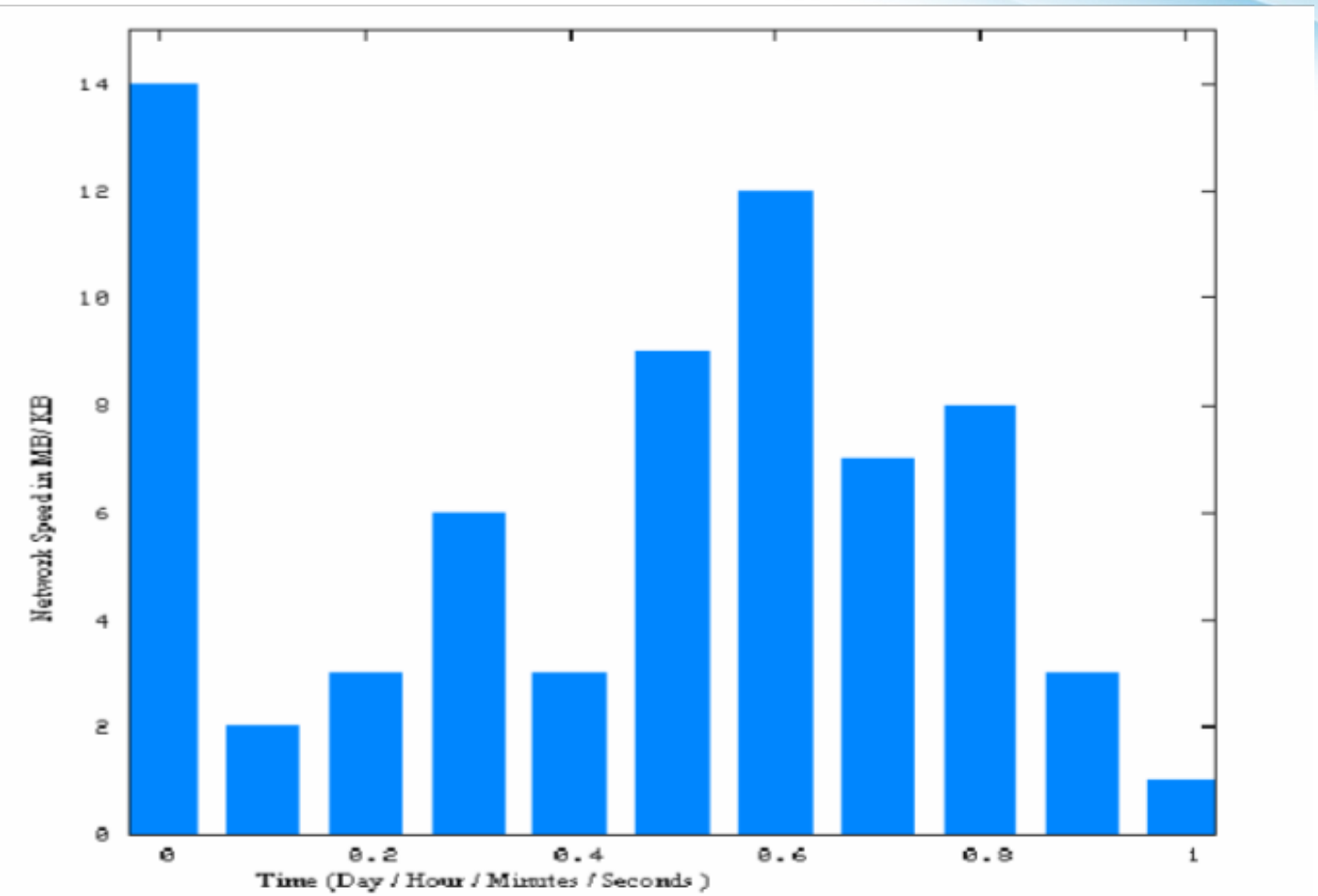
Features

- A Gigantic Repository containing download-able entities.
- Availability of the download-able URL are constantly checked and updated.
- Maintains a huge database of all available URL of resources such as Software / Study Materials in one place.
- Frequently update the repository
- Frequency of update can be customized.

Contd ..

- Download can be initiated either by manually or automatically.
- Status of download can be updated
- Download speed can be viewed in a graphical form

Graphical View of Download Speed of a File



Contd ..

- Automating the download of file(s)
- Intimating the User through E-Mail or some alarm, once download is complete

Table Schema of World Repo

The repository contains ,

- Name of the website from which the active link was crawled,
- Category of the domain, say it may be related to Engineering or Medical etc ,
- Meta-Data about the link say Description,
- Topics,
- Size,
- Active download link.

Customizing World Repo

- Notifying the Users about the updates through E-Mail.
- Notifying the users about the dead links
- Once this search engine is made to run in a regular interval it can be scaled up to crawl more and more websites

Making World Repo Narrower...

- Making World Repo, a bit narrower.
- Example ,using World Repo to get the price of a product

Implementation Details

- Crawling and Database Programming
- Crontab
- re
- pycha
- Smtplib
- Enhancing a Search Engine for Developer Community

Crawling and Database Programming

- Done by using urllib module in python
- f = urllib.urlopen(URL)

  Create file-like object that allows you to read the identified resource.
- Reads the identified resource and store it as a local file
- For Database programming, we use psycopg2 module in python

Crontab

- Crontab is a config file that specifies the shell commands to run periodically on a given schedule.
- Linux Utility
- Example:

Example:

Run the script every day at 12:00

```
0 12 * * *  /home/user_name/script.sh  > /dev/null 2 >&1
```

re

- re - regular expression.
- converting raw data to processed data .
- Done by Pattern matching.
- Achieved by using re module in python.

Pycha

- A Python package for drawing charts using the Cairo library.
- To graphically display the download statistics.

smtplib

- for notifying users through E-Mail regarding
  - repository updates
  - download status

Further Enhancements

- Give the user direct access to the tools he wishes to download.
- Although many Linux distributions provide such repositories, they are confined to the tools developed in that environment / operating system.
- An interface for the developer to find any software in the world developed across different platforms at one place.
- Support for Regional Indian Languages through Localization.

Questions

Thank You