**WEB CRAWLER**

**Flow Diagram:**

URLs

Download the webpage

Database

File Storage

1. Fetch Text data, Images (Maps), if constraint matches.

2. Send for storage.

**Web Crawler**

**Storage**

Constraint Based data Extraction

Ordered storage of fetched data

Scheduler

Hyperlinks/Sub URL information

Size of the Web page, Data Collection time

Excel Sheet

**Internet**

Queue

Hyperlinks/

subURLs

U

R

L

1. Receives seed URLs from user.

2. Pass URL for downloading website.

3. Remove URL from the list after passing.

4. Add Hyperlinks /subURLs to URL list send by queue.

1. Download the webpage of received URL/hyperlinks from www.

1. Get user Constraints.

2. Fetch information /hyperlinks based on constraints.

3. *Adaptive A\* search algorithm* visits the links until reach the goal webpage.

1. Queue adds Hyperlinks / subURLs to scheduler by following FIFO operation.

No of Images / Maps

1. MySQL DBMS manages data storage

1. Data stored under a specific column name given by the user.

**Procedure:**

1. Web crawler gets multiple URLs as input and adds these URLs to scheduler.
2. Scheduler passes URL for downloading whole webpage from the Internet, it may contain text data, images, hyperlinks, sub URLs etc.,
3. User has to mention which data is needed to fetch. If the requested data is in the form of Text or images, then it will be directly fetched and send to database / other file storage.
4. In database, crawled data will be stored in an ordered manner, it also find the number of images, size of the web pages and data collection time. In excel sheet data will be stored under a specific column name.
5. If collected information is in the form of hyperlinks / subURLs, then these are added to scheduler via queue. Now scheduler contains unvisited URL. Adaptive A\* Search algorithm is used to find the shortest path of the goal webpage using current link. It also updates relevancy value of the page using A\* Search and Best First Search algorithms.

Repeat steps from 2 to 4 until the scheduler is empty.

1. The contents of the webpage will be fetched in a particular interval based on the website changes and dynamically updated in a database/excel sheet automatically.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.No** | **Procedure** | **Explanation** | **Data Flow** |
| **1** | Web crawler gets multiple URLs as input and adds these URLs to scheduler. | 1. Web Crawler receives list of seed URLs as input. 2. Added URLs to scheduler. 3. Remove URL from the URL list after processing. 4. Hyperlinks/Sub URLs in a queue added to scheduler. | URL  Seed URLs  Scheduler  (Remove URL after sent)  Download webpage |
| **2** | Scheduler passes URL for downloading whole webpage from Internet, it may contain text data, images, hyperlinks, sub URLs etc., | 1. Download the corresponding webpage. 2. Webpage contains different type of information such as data, images, links etc., 3. Extract entire information of the website from the internet. | Website of inputted URL with text, images & hyperlink  Internet  U  R  L  Download  the  Webpage |
| **3** | User has to mention which data is needed to fetch. If the requested data is in the form of Text or images, then it will be directly fetched and send to database / other file storage. | 1. Get which information user need to fetch. 2. Check whether the webpage contains relevant information or not. 3. If content matches fetch text and images immediately and pass it to storage. 4. Otherwise extract any links contained in it. | Text and Image  User constraints  Constraint Based data Extraction  Fetch relevant information  Storage |
| **4** | In ***database***, crawled data will be stored in an ordered manner, it also find the number of images, size of the web pages and data collection time. In ***excel sheet***, data will be stored under a specific column name. | Database maintains the following information,   1. Ordered arrangement of text and images. 2. Number of Images. 3. Size of the webpage. 4. Data Collection time.   Excel sheet maintains data with specified column name given by the user. | Database  Ordered arrangement of text and images  Number of Images  Size of the webpage  Data Collection time  File Storage  Excel Sheet  (.csv format) |
| **5** | If collected information is in the form of hyperlinks / subURLs, then these are added to scheduler via queue. Now scheduler contains unvisited URL. Adaptive A\* Search algorithm is used to find the shortest path of the goal webpage using current link. It also updates relevancy value of the page using A\* Search and Best First Search algorithms.  Repeat steps from 2 to 5 until the scheduler is empty. | 1. For finding most relevant data, hyperlinks and subURLs are stored in a queue. 2. ***Adaptive A\* search algorithm*** visits the links until reach the goal webpage. 3. It uses A\* Search and Best First Search algorithms. 4. In BFS relevancy calculation is done for each link and the most relevant link, such as one with the highest relevancy value, is fetched from the frontier. 5. A\* Search uses BFS and find the search time. 6. And these links added to the scheduler. 7. URLs are processed, until return the most relevant page. | Hyperlinks/  subURLs  Hyperlinks/  subURLs  Constraint Based data Extraction  Scheduler  Queue |
| **6** | The contents of the webpage will be fetched in a particular interval based on the website changes and dynamically updated in a database/excel sheet automatically. | 1. Fix the time interval for each website, to continue this process automatically and dynamically. | Updates information dynamically & automatically for some fixed interval of time  Web Crawler  Storage |