**SOFTWARE AND HARDWARE REQUIREMENTS**

**SOFTWARE REQUIREMENT: -**

Front End/GUI Tool : PYCHARM

Operating System : Windows Family

Language : Python

Application : Windows Application

Back End : Mysql

**HARDWARE REQUIREMENT: -**

Processor : Pentium dual core

RAM : 4 GB

Hard Disk Drive : 500 GB

**EXISTING SYSTEM**

The existing system is more time consuming process because user can search the keyword after search complete existing system will list out no of URL based on keyword. User have to extract each and every Url manually then only the can get complete information. Existing system does not have any automatic Url information extraction Technique. It will not help to user get information in time with effective manner. The main drawbacks of the existing system can be summarized as follows:

**3.2 DISADVANTAGES**

* Time consuming process
* User have to extract each and every URL manually
* Difficult to know information
* Not a user Friendly process

**PROPOSED SYSTEM**

The drawbacks, which are faced during existing system, can be eradicated by using the proposed application. The main objective of the proposed system is to provide a user-friendly application so that user can get search information easily. Proposed application helps user can enter any type of keyword proposed system effectively extracts the content form URL list with the help of html parsing. After extract the content automatically content will be summarized based on user search keyword this information display in single page with effective manner.

**3.4 ADVANTAGES**

* Less Time consumption and makes the process of search more easy.
* proposed system is to provide a user-friendly application
* High level security.
* Large volumes of data can be stored with ease.
* Proposed apply Automatic information extraction and summarization technique.

**Modules**

1. **URL collection and upload**
2. **Information Search**
3. **URL mapping**
4. **Preprocessing**
5. **Content Extraction**
6. **Text summarization**

**URL collection and upload**

This module is mainly based on admin. System will check the admin user name and password for authentication. After the verification for authorization the admin can be able to enter complete collection of URL dataset list .This complete URL list will be maintain separate table

**Information Search**

Using this module user can enter any type of keyword this information goes to search process unit.

**URL mapping**

This is the important process based on user keyword this module automatically checks with data set URL list. Finally with the help of mapping algorithm it will map corresponding URL and display and store in array list effectively.

**Preprocessing**

This is most important process in data mining before process the document the given input document is processed for removing redundancies, inconsistencies, separate words, stemming and documents are prepared for next step.

**Content Extraction**

This module effectively extracts the content form URL list with the help of html parsing algorithm. This content will be maintained separately for content grouping process.

**Text summarization**

This is the final module after extract the content automatically content will be summarization based on user search keyword this information display in single page. so that user no need to visit each and every URL link separately .

**DATA FLOW DIAGRAM**

DFD depict hoe data interact with the system. DFD are extremely useful in modeling many aspects of a business function because they systematically subdivide a task into basic parts, helping the analyst understand the system that they trying to model.

A data flow diagram models a system by using external entities from which data flow to a process which transmission the data and creates output data which goes to other processes on external entities of files. Data may also flow to process as inputs.

The symbols appearing in the DFD has been explained below:

- Represents a process

- Which shows data flow

- Designation of the data

- Shows Data source

**LEVEL 0**

Receive content information

Login and search

User

Admin

Add URL Dataset /View

Login Process

Level l: Admin

URL Dataset entry

If login successful

User id and password

Login table

**Admin**

URL table

Store URL details

Get details

Report

**Level 2:**

Enter id

User register

User

Search table

Password, etc

URL table

Enter Keyword

If login successful

Retrieve for authentication

Data summarization

**System Flow Diagram**

**Search Keyword**

URL MAPPING

**URL data set**

**Processing**

**Data Extraction**



User

**Html parsing**

**Data Summarization**

**Er Diagram:**

URL

Registration

Search

URL

**15. TABLE DESIGN**

**URL dataset**

**Primary Key: Url id**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Data type** | **constraints** | **Description** |
| Url Id | Numeric(18,0) | Not Null | It Contains Url Id |
| URL name | NVarchar(Max) | Not Null | It Contains URL name |
| Attribute | NVarchar(Max) | Not Null | It Contains Attribute |
| Value | Varchar(30) | Not Null | It Contains Value |
| Description | NVarchar(Max) | Not Null | It Contains Description |

Registration Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Data type** | **constraints** | **Description** |
| Name | Varchar(10) | Not Null | It Contains Name |
| Password | Varchar(10) | Not Null | It Contains Password |
| DOB | Varchar(45) | Not Null | It Contains DOB |
| Gender | Varchar(30) | Not Null | It Contains Gender |
| Address | Varchar(20) | Not Null | It Contains Address |
| Country | NVarchar(Max) | Not Null | It Contains Country |
| Mobile Number | Int(10) | Not Null | It Contains Mobile Number |
| Email id | NVarchar(Max) | Not Null | It Contains Email id |

**Data summarization Table:**

foreign key: Url id

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Data type** | **constraints** | **Description** |
| search | Numeric(18,0) | Not Null | It Contains Data id |
| Url id | NVarchar(Max) | Not Null | It Contains Data Url |
| Title | NVarchar(Max) | Not Null | It Contains Data Title |
| Content | NVarchar(Max) | Not Null | It Contains Data Content |
| Tags | NVarchar(Max) | Not Null | It Contains Data Tags |
| Para | NVarchar(Max) | Not Null | It Contains Data Para |
| Links | NVarchar(Max) | Not Null | It Contains Data Links |
| Extraction Time | Date time | Not Null | It Contains Extraction Time |

Form Design











