IQRA UNIVERSITY IU

IWS

Smart Searching

INTELLIGENT WEB SPIDER FOR EFFICIENT SEARCH ENGINE (IWS)

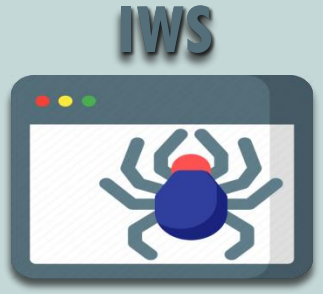
PRESENTED BY:

Abdullah Rather (35472)
Bilal Ahmed (35708)
Muhammad Hassaan (35387)
Syed Farid Uddin (35485)

SUPERVISOR:

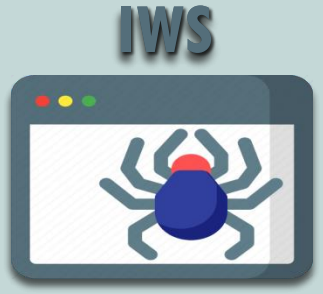
Sir Israr Ali

INTRODUCTION



What is IWS?

- IWS is an Internet Web Crawler along with a Search Engine
- It is a combination of Desktop & Web based Applications
- It performs crawling on specified websites to generate keywords and store them in hierarchal clustered Database.
- This application allows a user to search on its Search Engine and provides a list of URLs containing the searched keyword.
- It helps user identify website containing most information related to their searched keyword.
- It maintains the frequency of keywords generated from different URLs and lists them in hierarchal format according to the frequency/repetition of a word on URLs top to bottom.



PROBLEM SUMMARY

- The databases for search engines are populated using automated tools called web crawler, these tools are used to see contents of any site. But these tools are costly and inaccessible to students and small level professionals.
- These tools don't provide complete access & authentication to users, e.g., User can't define level of depth for searching or specific domain level searching / keyword generation.
- The keywords generated by such tools are very large in number and has huge repetition of keywords.
- Usually storing keywords requires extensive storage as keywords are generated in bulk
- There is a requirement to design a web crawler that should generate keywords while conserving space, it should be able to store generated keywords without affecting the performance of Search Engine.

STATE OF ART

History:

Literature review is conducted before the development of the proposed system 'Intelligent Web Spider for Efficient Search Engine'. Many crawling software were in the domain of this survey. Some of them are listed here with brief description.

- **Trellian SiteSpider**

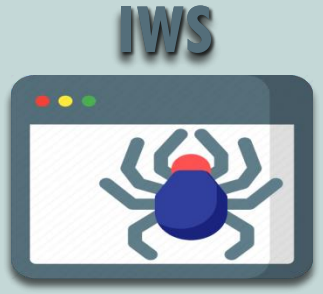
  Trellian SiteSpider can help to extract valuable data from any website

- **Icegiant Data Extractor**

  MP3 Extractor is a state-of-the-art internet MP3 download manager.

- **Inspyder Site Map Creator**

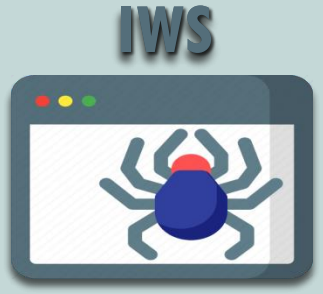
  Inspyder Sitemap Creator is a web-crawling XML sitemap file generator.



STATE OF ART

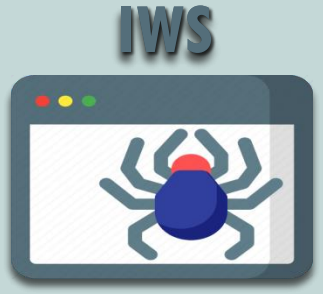
What is being done now?

Our objective is to investigate ways to offer raw material to search engines in order to improve their efficiency and provide better results while preserving the smallest feasible footprint and storage space, all while keeping the significance of efficient searching in mind.



STATE OF ART

- Search engine are state of the art tools in use now a days
- Our solution is based on state of the art tools and technologies
- The design is user friendly and comply with all User Interface and User Experience (UI/UX) principles
- Overall efficiency of the system is overwhelming which may be seen at each step of the application flow



PROPOSED SOLUTION

Mission Statement:

While keeping the importance of efficient searching in mind, our mission is to explore ways to provide raw material to the search engines in order to enhance their efficacy and generate better results while maintaining lowest possible footprint and storage size.

PROPOSED SOLUTION

Scope of the Project:

An Intelligent Web Crawler that should supports Search Engine in performing efficient searches by providing purified data captured from webpages and processed through Linguistic Algorithm in order to simplify Keywords for searching and their efficient storage.

PROPOSED SOLUTION

Objectives:

- To identify a viable linguistic algorithm that could simplify and transform keywords so that storage could be minimized
- To do the research in exploring linguistic algorithm that is best suited for our needs
- To explore efficient storage mechanism for keywords in database clusters

PROPOSED SOLUTION

Benefits:

- Website navigation to fetch its contents
- Collect keywords from any website around the world
- Apply linguistic algorithm to filter the keywords
- Test links of the website to identify its valid structure/hierarchy.
- While crawling through the website, it can also identify bad links (the links which are invalid or are without target pages)
- It may search for copyright violations by filtering keywords.
- Clustered based keyword storage for efficient retrieval
- Maintain ratings of keywords

PROPOSED SOLUTION

Outcomes/Final Product

- Fetch keywords from the contents of a website
- Make a list of URLs after crawling and make a web of URLs for subsequent keyword fetching
- Simplify keywords by applying linguistic algorithm
- Store keywords
- Maintain frequency of keywords for better searching

PROPOSED SOLUTION

Features:

- Fetch Keywords from website.
- Separate Non important keywords.
- Make list of URLs by Crawling and make a web of URLs for subsequent Keyword Fetching.
- Apply Stemming Algorithm to simplify keywords and reduce storage requirements.
- Maintain site-wise Frequency of Keywords.
- Manage keywords in manageable Clusters based on their categories.
- Controlled Crawling option.
- Apply Search from keywords Database.

PROPOSED SOLUTION

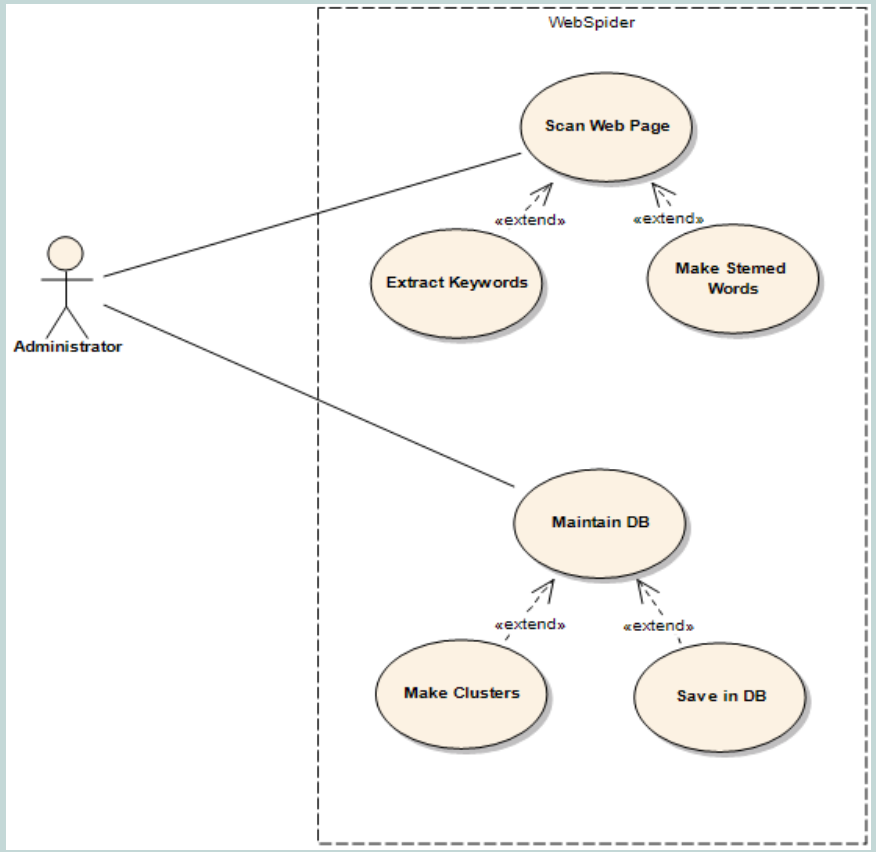
Implementation issues & Challenges:

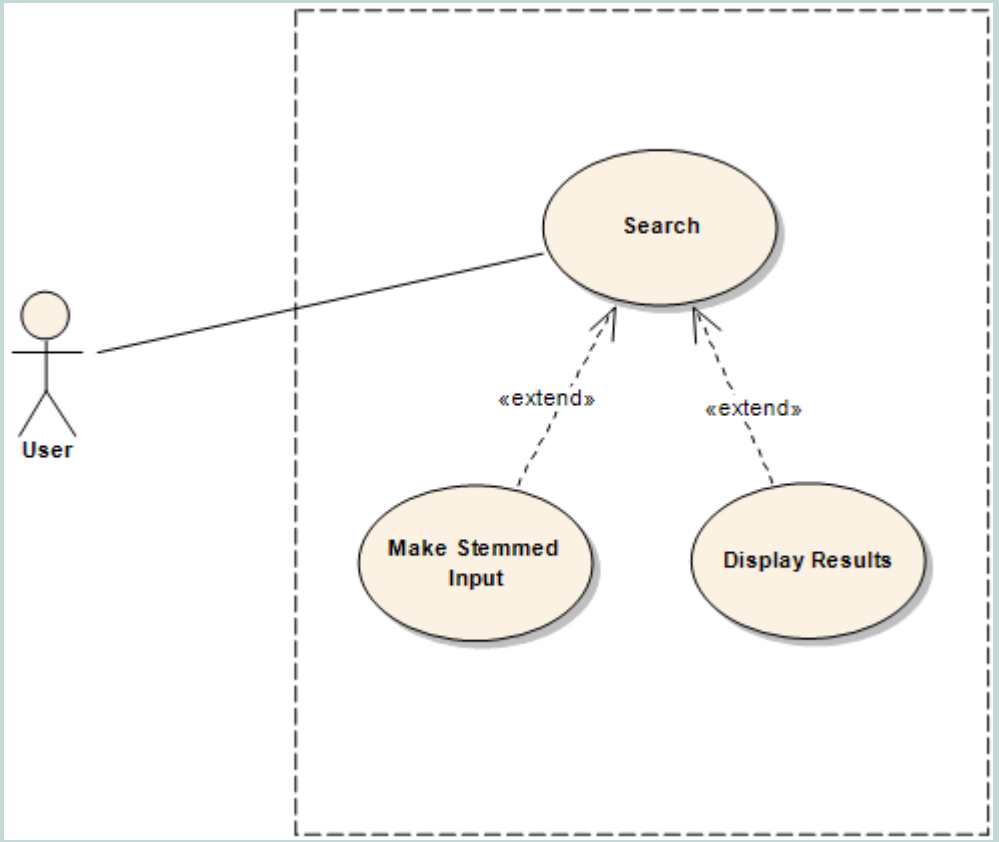
- Crawlers are costly tools to be used for research purposes and generally inaccessible.
- Efficient storage optimization of generated keywords is a big challenge.
- Usually, Crawlers generate keywords in huge numbers, and they do not provide the flexibility to define level of depth for searching or specific domain level searching / keyword generation.
- Fetching keywords from secure websites (i-e protected by SSL) is a challenge
- Identification and coding of Linguistic Algorithm
- Integration of overall system

IWS

Smart Searching

USE CASE/PRIMARY AUDIENCE





TARGET MARKET

IWS can be used in data intensive markets where data set is big and searching on such Big Data takes time. Some of the examples are as follows:

- Business industry.
- Defense.
- Electronic Media.
- Universities.
- Government and Private offices.

IWS

Smart Searching

PLATFORM/TECHNOLOGIES

- Java
- PHP
- MySQL
- Apache NetBeans
- WampServer

WHY IS OUR PROJECT UNIQUE?

- Our project is unique because it generates keywords and apply linguistics algorithm to purify them and simplify their storage to conserve space for quick and efficient searching.
- As our project covers end to end cycle used by search engines including its own crawler, keyword generation, their storage in our defined data clusters and a web-based search engine component that search from our own generated keywords. This whole coverage makes our project Unique compared to other similar solutions.

DIVISION OF WORK

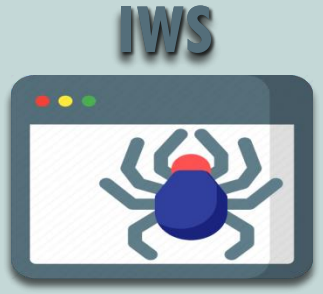
We are a group of four Individuals with different skill level and mind sets, who worked together to find solution on a single problem.

The Group members and their responsibilities are as follows:

- **Abdullah Rather** (Project Lead, Design, Development, QA and Documentation)
- **Bilal Ahmed** (QA Engineer)
- **Muhammad Hassaan** (Back-end Developer)
- **Syed Farid Uddin** (Front-end Developer)

And the Project concluded under the supervision of :

- **Sir Israr Ali** (Assistant Professor)



CONCLUSION

- Keywords generated by the crawler are based on the text it captures from the website. There are some useless words that should be skipped while adding the keywords. We applied this logic by adding skipping words in a text file with our project. This text file (we are calling it as “Stop List”) currently is based on the words we manually populate in the text file.
- In future, we intend to automate the Stop List based on some Artificial Intelligence (AI) Techniques so that more and more accurate results should be populated and crawler should learn with the passage of time to filter useless words without human intervention.

THANKYOU !