DEVELOPMENT OF SEARCH ENGINE

PRESENTED BY:-

Khushboo Kumari

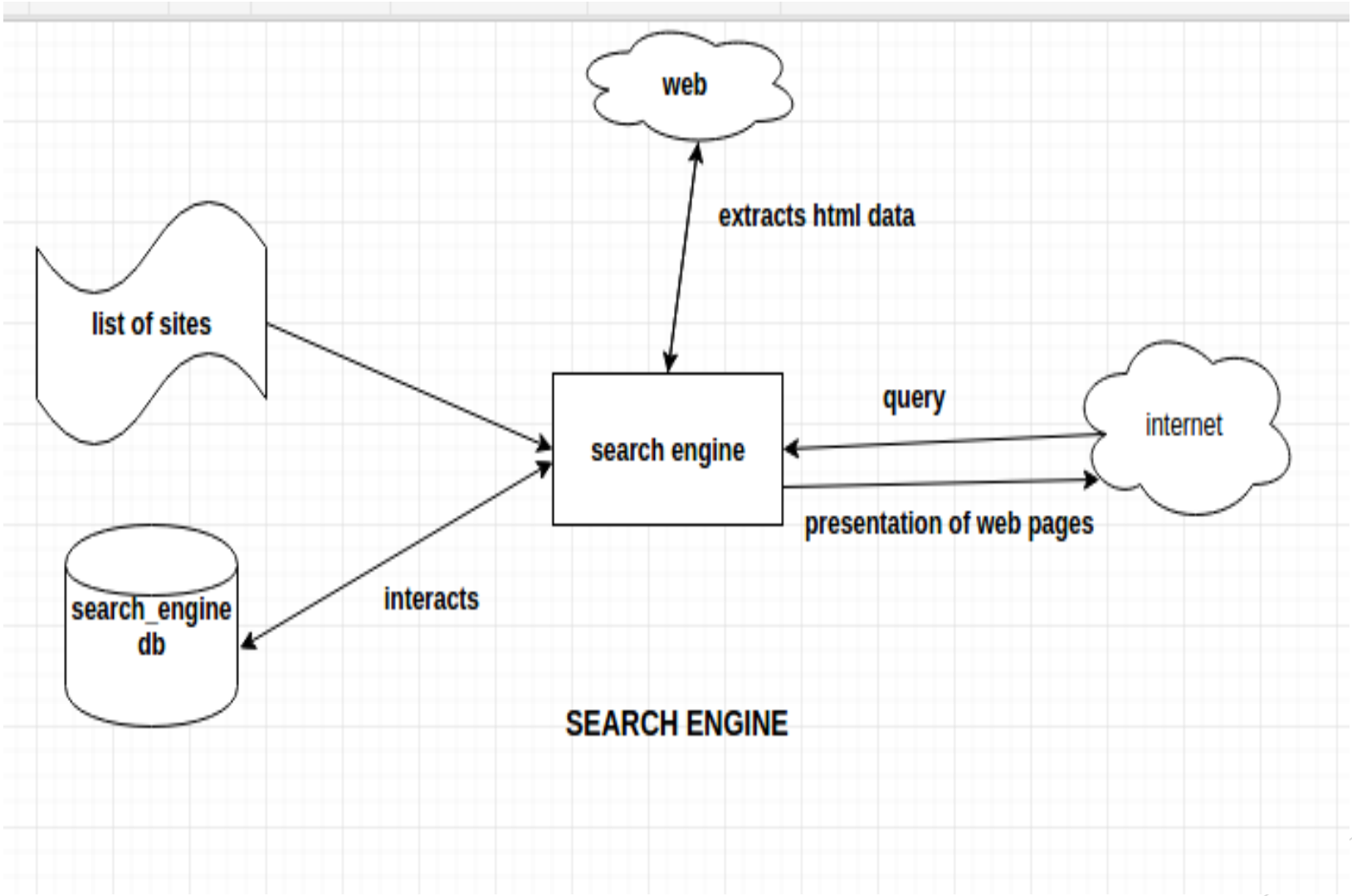
Neeraj Kumar

Shweta Kumari

Purpose and Features:-

- It has been designed to help people find information stored on one or more than one sites.
- A search engine, that search web links for specified keywords and returns a list of web links and some code snippets where keywords are found.
- Scalable
- Ranking can be modified using Weight Configuration Table
- Advertisement Free unlike Google Custom Search Engine

Overall Architectural Diagram



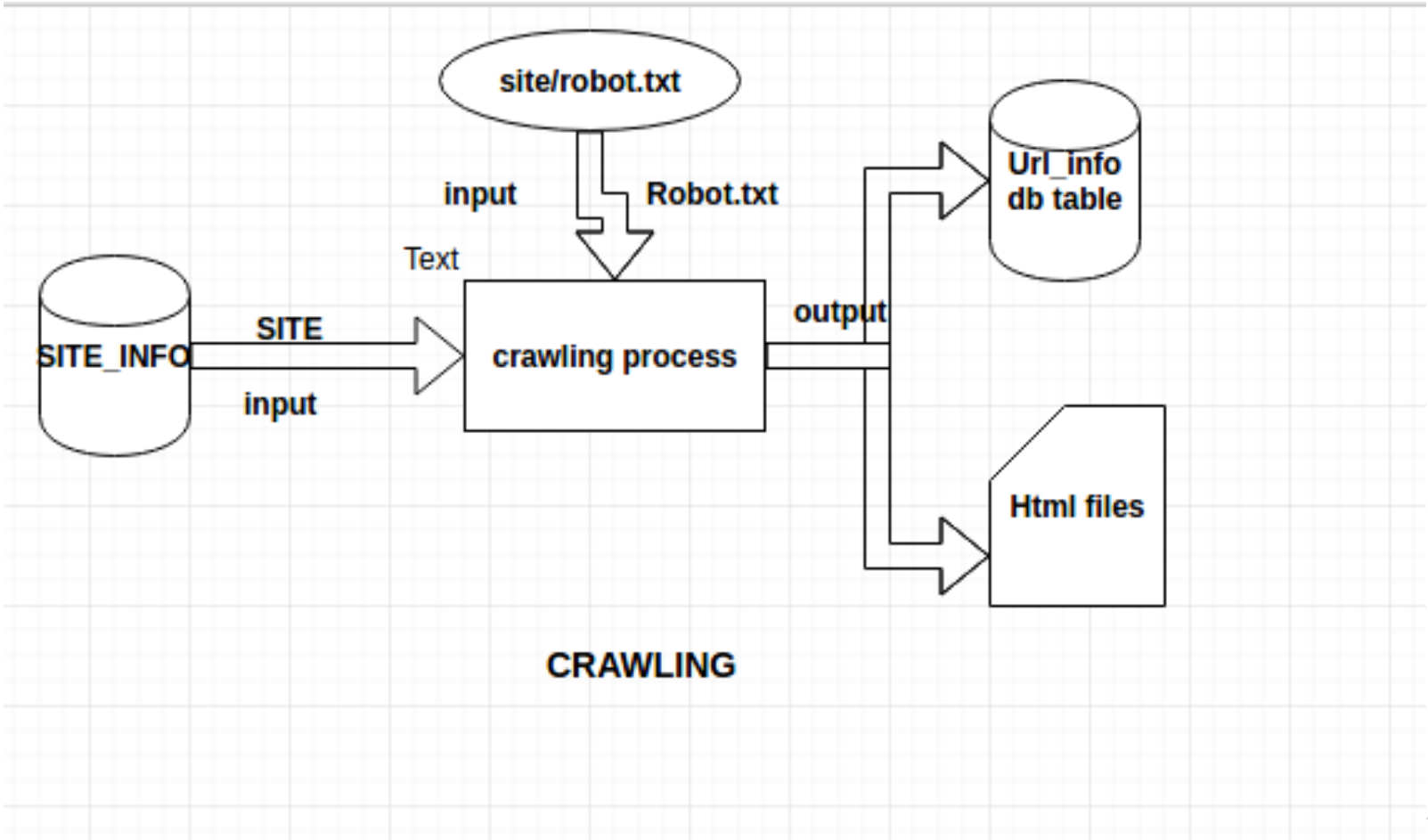
Modules

- There are mainly four modules
  - Crawling
  - Indexing
  - Ranking
  - Presentation

Crawling

- Crawlers look at web-pages and follow links on those pages.
- They go from link to link and bring data about those web-pages.
- It takes the input as list of site links and crawl all the urls and stores in a database table.
- Input:-List of Site urls
- Output:- URL INFO table corresponding to all the urls in that site and html files

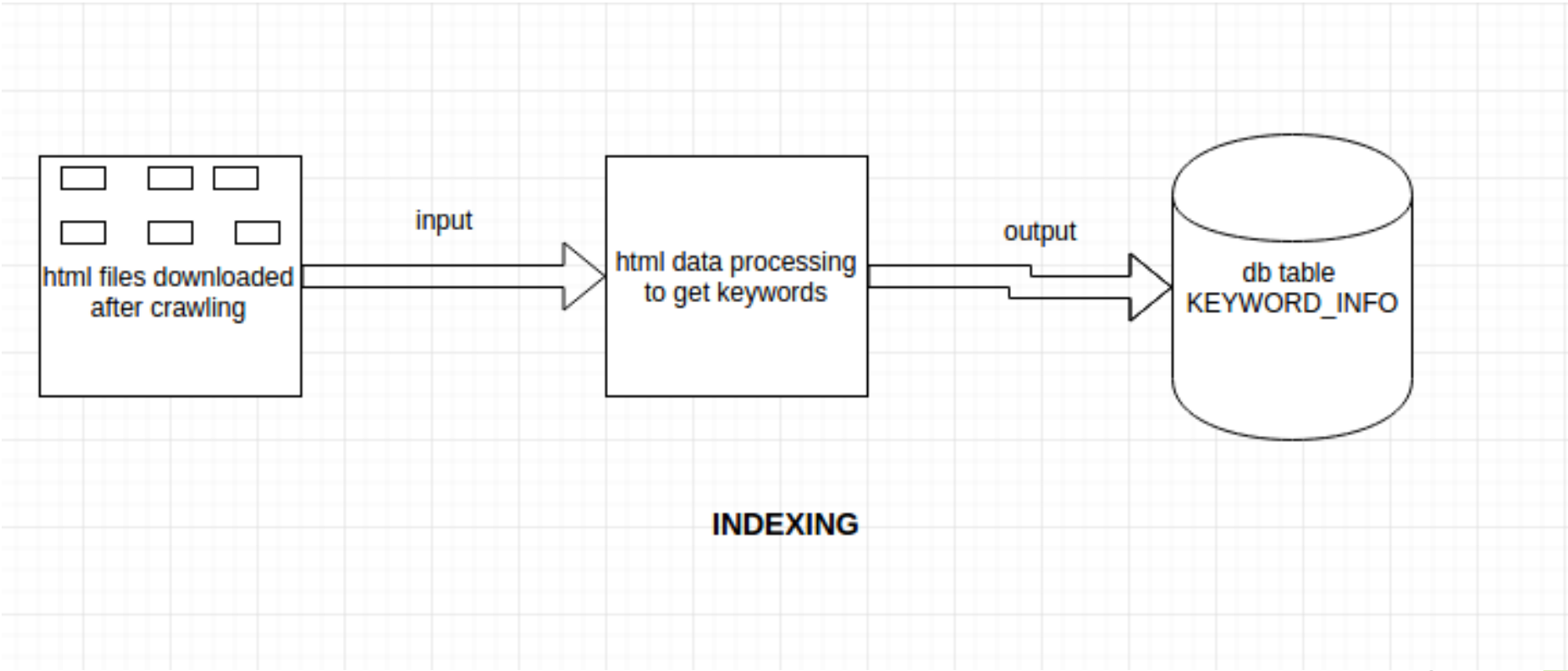
Crawling Architecture



Indexing

- Includes various methods for indexing the contents.
- Html text, title, meta description/keywords has been used for extracting the keywords .
- Keywords are stored in the table KEYWORD INFO.
- Occurrence of keywords has been also taken.
- Input: html data downloaded during crawling
- Output: KEYWORD INFO table

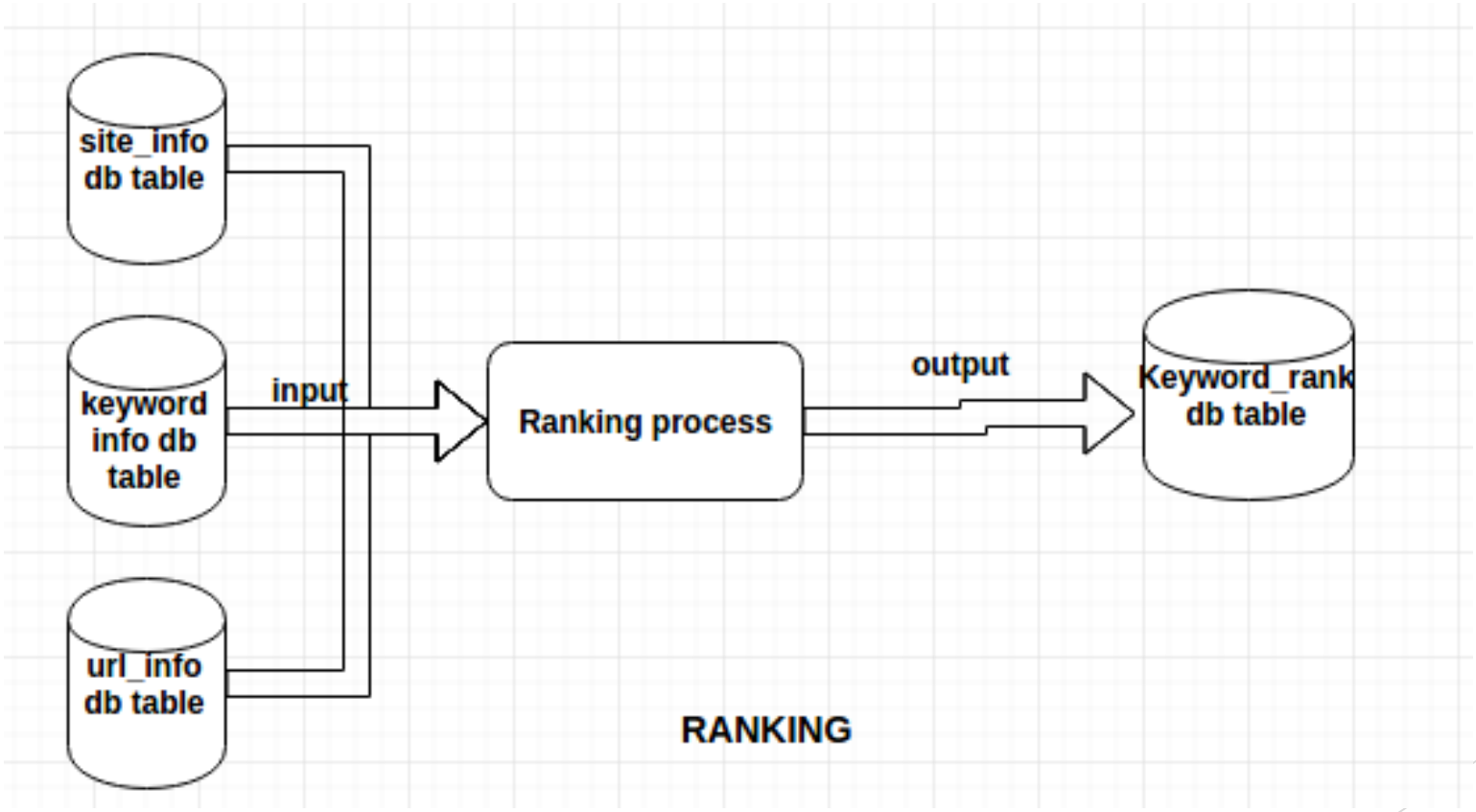
Indexing Architecture



Ranking

- Relevant urls containing query keywords are ranked using different factors.
- Find the urls in which query keywords are present and then based on the weight of url, site and occurrence weight, urls are ranked.
- Site having good reputation, more number of viewers, good links and other factors are given more weight-age compared to others sites and same goes for urls.
- Input: html data downloaded during crawling –
- Output: KEYWORD INFO and KEYWORD RANK table

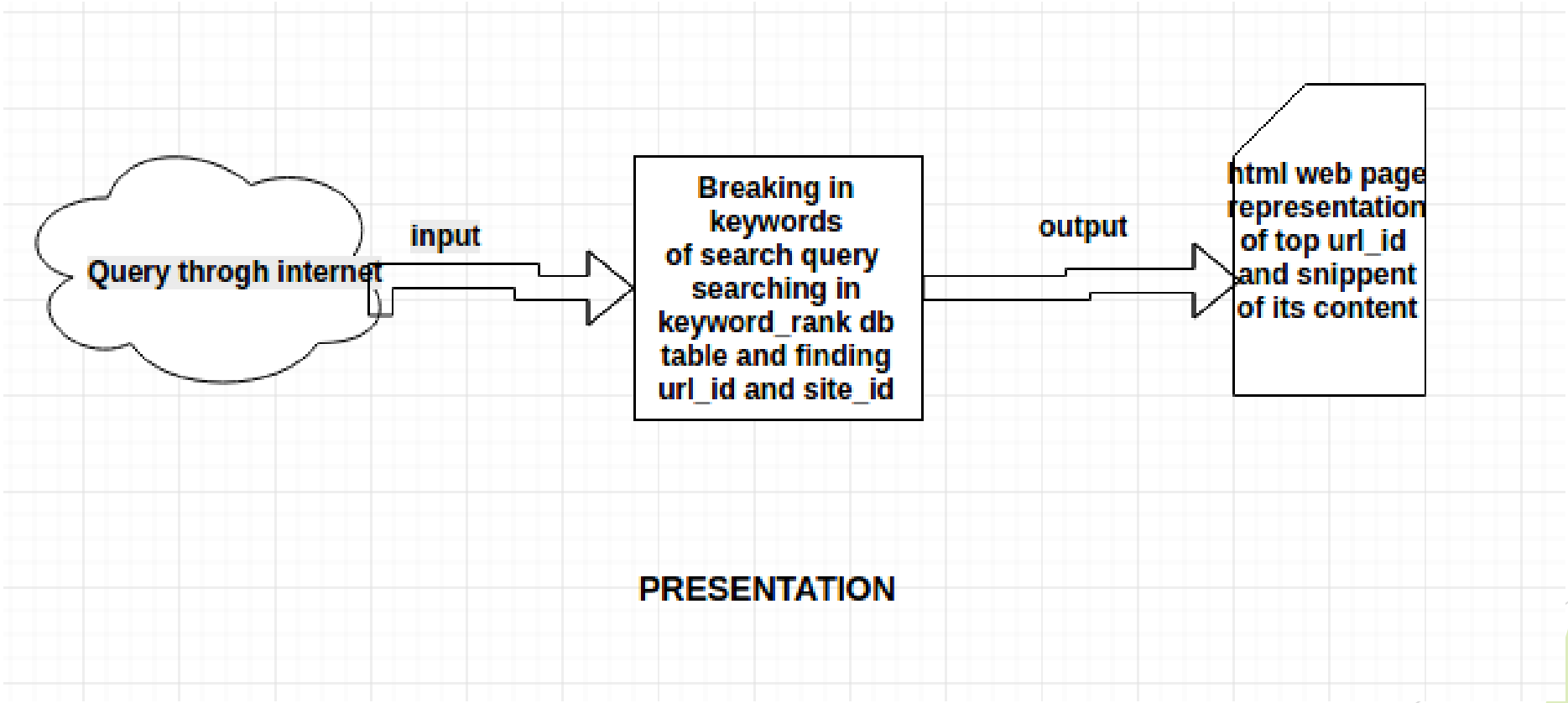
Ranking Architecture



Presentation

- Displaying the links with title and some code snippet based on the result obtained after ranking.
- There will be a html page containing search form which will take input from the user and based on the back-end processing, it will display the result.
- Input: Query from user
- Output: Search result

Presentation Architecture



Ranking Parameters

- Domain age
- Keyword in body
- Meta desc
- Meta keyword
- Title
- H1 heading
- H2 heading
- http/https(security)
- Url-name
- Url-site

Scalability:-

- It will provide scalability
- Have the ability to run different modules at different servers.
- Therefore large number of websites can be crawled and stored in the database.
- Implemented by giving an attribute SERVER_IP in the database table site_info and FileLocation in Keyword Rank Table
- Instead of sqlite3, MYSQL language can be used.

To do:

- Edit Distance Algorithm can be used
- Presentation of text can be improved
- Ranking Parameters to be implemented
  - Number Of viewers
  - Bounce Rate
  - Inbound Limit
  - Content Quality Factors
    - rel=”no follow" is present or not in the anchor tag
    - Alt is present in image tag or no

THANK YOU