**Journal of Image Processing and Intelligent Remote Sensing**
**ISSN 2815-0953**
Vol: 02, No. 03, April-May 2022
http://journal.hmjournals.com/index.php/JIPIRS
**DOI:** https://doi.org/10.55529/jipirs.23.10.13

# A Model for Gujarati News Search Engine by Link Builder and Web Crawler Algorithms

**Hemang Desai[1*], Birajkumar V. Patel[2]**

*[1*]Vimal Tormal Poddar BCA College, Surat, Gujarat, India.*
*[2*]P.G. Department of Computer Science S.P. University, Vallabh Vidhyanagar, Surat, Gujarat, India.*

**Received:** 26 December 2021 **Accepted:** 22 March 2022 **Published:** 18 April 2022

***Abstract: The web serves a paradoxical role in language learning. Providing a fertile ground for neologisms and redefinitions, the web incubates a growing translation gap between one's native tongue and his or her target foreign language. However, the web can also be used to bridge languages even as they change, because it contains a vast and ever-expanding set of human-translated web pages. Search engines that mine high-quality human translations can unlock the web's pedagogical potential, and form the basis for next-generation language learning tools. So, it has become essential to develop crawlers in such a way that it can help in local language optimization with topic-search. Based on study of many research papers motivated to do research in Gujarati Search Engine. Moreover, proposed a new model for it.***

***Keywords: Web crawler, Link Builder***

## 1. INTRODUCTION

The traditional search engine works with information retrieval and out putted those to the user whether it is relevant or not. In the literature, research focused on topic-oriented search for the user in specific format that can be document, news, picture, video etc. [1] [2].

In this search, focused on Guajarati newspapers (e-papers) which are published on daily basis on the web. So, the search engine behaves like topic-search for the Guajarati newspapers which are published every day on the web.

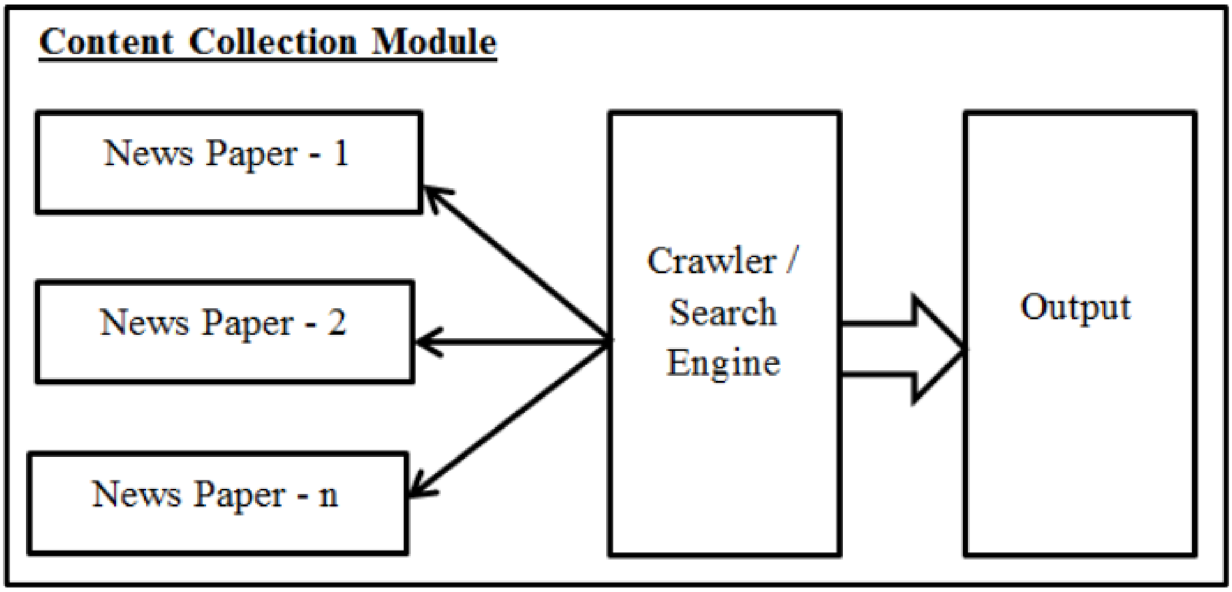

Figure 1: Architecture of Proposed Search Engine Model

Above Figure 1 depicts the architecture of the proposed search engine. This architecture is divided into two different parts such as availability of electronic Gujarati newspapers on web & Crawler. Here, link builder sub module of crawler works for relevant topic search. Here, Link builder is used to build relevant link based on the topic searched by the user in the search engine.

Here, developed a search engine to search news in Gujarati language from selected websites of Gujarati news providers (e-papers).

The proposed model includes newly developed Algorithms:

**Proposed Algorithms and Flowcharts**

**Algorithm for Link Builder:**

```
Step 1 : Start
Step 2 : Read  Gujarati Newspaper websites from Datasets.
Step 3: Visit All links category wise from the files,
Step 4: Make a request to the URL one by one .
Step 6 : Fetch all response.
Step 7 : Extract all the <a> anchor tags.
Step 8 : All tags save in the database with a unique page hash key.
Step 9 : Keep a flag as "Unvisited URL" for the specific URL.
Step 10: Stop
```

### Flowchart for Link Builder

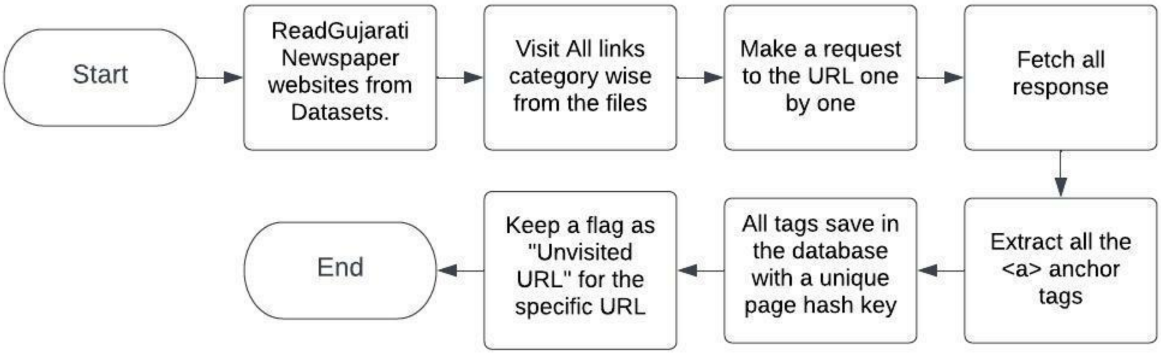
Start
ReadGujarati Newspaper websites from Datasets.
Visit All links category wise from the files
Make a request to the URL one by one
Fetch all response
Extract all the <a> anchor tags
All tags save in the database with a unique page hash key
Keep a flag as "Unvisited URL" for the specific URL
End

Figure 2: Proposed Flowchart for Link Builder

**Algorithm for Web Crawler:**

Step 1: Start
Step 2: Fetch all the saved URLs(Unvisited URL) found in the database.
Step 3: Clean up the response HTML
Step 4: Clean all HTML tags
Step 5: Remove all Script, Style and inline CSS
Step 6: Save page content for database search query.
Step 7: Mark the flag of saved URL as Visited URL.
Step 8: Stop

### Flowchart for Web Crawler:

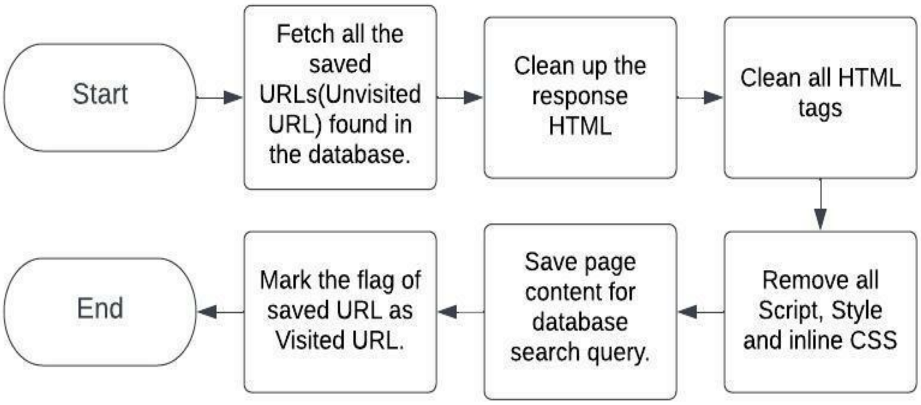
Start
Fetch all the saved URLs(Unvisited URL) found in the database.
Clean up the response HTML
Clean all HTML tags
Remove all Script, Style and inline CSS
Save page content for database search query.
Mark the flag of saved URL as Visited URL.
End

Figure 3: Proposed Flowchart for Web Crawler

## 2. CONCLUSION

People are always searching content from the internet for their specific requirements such as News, Videos, Images, Blogs etc. Here, in this research work, studied research papers from the web and derived a conclusion that it is necessary to have one specific model for the Gujarati news search engines. Hence, in this research paper, developed a quantifiable and sophisticated Model for the Gujarati News search engines. To develop this research model, designed, developed, and implemented two algorithms also for Gujarati News Link Builder and Web Crawler. So, this research will be greatly useful for internet users to search news in Gujarati language.

## 3. REFERENCES

1. Mahale, V. V., Dhande, M. T., & Pandit, A. V. (2018, August). Advanced web crawler for deep web interface using binary vector & page rank. In 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on (pp. 500-503). IEEE.
2. Chhabra, S., Mittal, R., & Sarkar, D. (2016, August). Inducing factors for search engine optimization techniques: A comparative analysis. In 2016 1st India International Conference on Information Processing (IICIP) (pp. 1-4). IEEE.