Web Crawler: Sequential, Parallel, and Distributed Implementation

Vanja Antonovic
University of Primorska – Famnit
Koper, Slovenia
89211069@student.upr.si

Abstract

This paper outlines the development of a domain-restricted web crawler in Java. The project was implemented in three execution modes: sequential, parallel (multithreaded), and distributed (MPJ Express). Each version was evaluated for scalability, performance, and accuracy in extracting and validating internal hyperlinks. The crawler identifies both working and broken links and generates summary reports. The paper includes technical design, code snippets, architectural considerations, challenges encountered, implementation limitations, benchmarking methodology, and potential future improvements.

1 Introduction

A web crawler is an automated program that systematically navigates through websites, extracts links, and analyzes page content. The objective of this project is to build such a system using Java in three progressively scalable implementations:

- A single-threaded sequential crawler using breadth-first search (BFS)
- A multithreaded crawler using Java's concurrency API
- A distributed version leveraging MPJ Express for parallelism across JVMs

The project is designed with modularity and extensibility in mind. Common components such as URL parsing, validation, and link extraction are shared between implementations, allowing experimentation and comparison across architectures.

1.1 Motivation

Web crawlers are foundational tools in data mining, search engines, and content aggregation. By implementing various versions, we aim to understand the performance tradeoffs in computational overhead, parallelization complexity, and fault tolerance.

2 Problem Definition

The crawler accepts an initial root URL and maximum depth. It recursively discovers internal hyperlinks, classifies them by status (valid or broken), and logs crawl paths. Requirements:

- Restrict crawl to the same domain
- Record source page, discovered links, and HTTP response codes
- Avoid visiting the same URL multiple times
- Handle invalid URLs and HTTP exceptions

The system must be benchmarked across different modes with identical inputs for fair comparison.

3 Sequential Implementation

3.1 Overview

The sequential implementation uses a basic BFS traversal pattern. It performs the crawl on a single thread and is suitable for smaller websites or initial testing. We use JSoup to parse HTML and extract anchor tags.

3.2 Core Data Structures

- **Queue:** `LinkedList<URLDepthPair>` – tracks unvisited URLs with depth
- **Set:** `HashSet<String>` – tracks visited URLs

3.3 Code Sample

Sequential BFS Web Crawler

```
Queue<URLDepthPair> queue = new LinkedList
    <>();
Set<String> visited = new HashSet<>();
queue.add(new URLDepthPair(root, 0));
while (!queue.isEmpty()) {
  URLDepthPair current = queue.poll();
  if (!visited.contains(current.url)) {
    visited.add(current.url);
    List<String> links = extractLinks(
        current.url);
    for (String link : links) {
      if (isInternal(link) && current.depth
          < MAX_DEPTH)
        queue.add(new URLDepthPair(link,
            current.depth + 1));
    }
  }
}
```

3.4 Advantages and Limitations

Pros: Simple to debug, low overhead.
Cons: Poor scalability, blocked by slow servers.

4 Parallel Implementation

4.1 Design Goals

The parallel version improves responsiveness and throughput by assigning URL extraction tasks to worker threads. The main challenge is ensuring thread-safe access to shared queues.

4.2 Implementation Strategy

- Thread pool with 4–8 threads
- `Collections.synchronizedSet` for visited tracking
- Synchronized queue for thread-safe polling

4.3 Code Sample

Threaded Worker

```
Runnable crawlerWorker = () -> {
  while (!queue.isEmpty()) {
    String url;
    synchronized (queue) {
      url = queue.poll();
    }
    if (url != null && !visited.contains(url
        )) {
      visited.add(url);
      List<String> links = extractLinks(url)
          ;
      synchronized(queue) {
        for (String link : links)
          if (!visited.contains(link)) queue
              .add(link);
      }
    }
  }
};
```

4.4 Analysis

Multithreading significantly reduced crawl time at depth 3, but performance gains diminish at higher depths due to contention and shared-state locking.

5 Distributed Implementation with MPJ Express

5.1 Approach

MPJ Express provides message-passing primitives across JVMs. One master process (rank 0) assigns URLs; worker ranks crawl and return results.

5.2 Code Sample

MPJ Communication

```
MPI.COMM_WORLD.Send(sendData, 0, sendData.
    length, MPI.OBJECT, 0, 99);
MPI.COMM_WORLD.Recv(recvData, 0, recvData.
    length, MPI.OBJECT, 0, 99);
```

5.3 Shortcomings

- No shared memory or global queue
- Requires manual balancing
- Poor fault tolerance

6 Architecture Overview

6.1 Component Structure

- **URLDepthPair** – container class holding a URL and its corresponding depth level.
- **SequentialCrawler, ParallelCrawler, DistributedCrawler** – main classes handling the crawling loop for each mode.
- **JSoup Parser** – external library used directly to fetch and parse links from HTML pages.
- **ThreadPoolExecutor** – used in the parallel crawler to manage thread concurrency and task submission.
- **MPI Communication** – used in distributed mode to send tasks and receive results across JVM ranks using MPJ Express.

6.2 Data Flow

(1) Input: Root URL, Max Depth
(2) Extraction → Queue → Deduplication → Validation
(3) Output: Reported to console or log file

7 Error Handling

Implemented with try-catch blocks. Failure logs distinguish between:

- **Transient:** Network issues (e.g., 503)
- **Permanent:** Malformed URL, unsupported protocol

8 Retry and Resilience

Although retries are not built-in, we propose the following:

- Use exponential backoff for 5xx errors
- Skip retry on 4xx errors
- Batch log failed retries

9 Testing Methodology

9.1 Environment

Tests were run on a quad-core system, crawling `https://www.famnit.upr.si`.

9.2 Test Scenarios

- Same root URL
- Depths: 1, 2, 3
- Output validated for uniqueness, status, and time

9.3 Metrics Collected

- Execution time (ms)
- Pages visited
- Broken vs working link ratio

10 Limitations

- No robots.txt parsing
- JSoup cannot render JavaScript-heavy content
- Distributed version lacks real-time coordination

11 Results

Table 1: Execution Times

Depth	Sequential	Parallel	Distributed
1	100ms	55ms	75ms
2	700ms	400ms	460ms
3	2600ms	1300ms	1540ms

12 Future Work

- Visual UI using WebSocket for live crawl visualization
- Priority-based queue (e.g., prioritize unvisited domains)
- Persistent logging via SQLite or CSV
- JavaScript rendering with Selenium integration
- robots.txt compliance layer
- Recoverable job partitioning for MPJ (e.g., heartbeat or state-file)

13 Conclusion

The crawler system was implemented progressively, improving from a basic sequential strategy to a fully distributed prototype. The project demonstrated:

- Importance of BFS for controlled crawling
- Value of parallel workers in reducing latency
- Complexity of distributed crawling with message passing

Performance improved consistently in multithreaded mode, while MPJ Express enabled scaling across multiple nodes. Though limited in JavaScript rendering and fault tolerance, the architecture sets a solid foundation for future improvements.

References

[1] MPJ Express, https://mpj-express.org/
[2] Vanja Antonovic, Web Crawler GitHub Repository, https://github.com/iSoulFeed1g/web-crawler

```
[INFO] [main] 2025-07-30 13:02:07 - Root offered to queue: true, visited added: true
[INFO] [Worker-3] 2025-07-30 13:02:07 - [Thread Worker-3] Crawling (Depth 0): https://www.famnit.upr.si
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/ (Depth 1)
[INFO] [Worker-4] 2025-07-30 13:02:08 - [Thread Worker-4] Crawling (Depth 1): https://www.famnit.upr.si/sl/
[INFO] [Worker-3] 2025-07-30 13:02:08 - Skipped (different domain): http://www.upr.si
[INFO] [Worker-3] 2025-07-30 13:02:08 - Skipped (different domain): http://www.upr.si
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/ (Depth 1)
[INFO] [Worker-2] 2025-07-30 13:02:08 - [Thread Worker-2] Crawling (Depth 1): https://www.famnit.upr.si/
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl (Depth 1)
[INFO] [Worker-6] 2025-07-30 13:02:08 - [Thread Worker-6] Crawling (Depth 1): https://www.famnit.upr.si/sl
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/en (Depth 1)
[INFO] [Worker-7] 2025-07-30 13:02:08 - [Thread Worker-7] Crawling (Depth 1): https://www.famnit.upr.si/en
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/izobrazevanje/dodiplomski-studij (Depth 1)
[INFO] [Worker-5] 2025-07-30 13:02:08 - [Thread Worker-5] Crawling (Depth 1): https://www.famnit.upr.si/sl/izobrazevanje/dodiplomski-studij
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/izobrazevanje/podiplomski-magistrski-studij (Depth 1)
[INFO] [Worker-8] 2025-07-30 13:02:08 - [Thread Worker-8] Crawling (Depth 1): https://www.famnit.upr.si/sl/izobrazevanje/podiplomski-magistrski-studij
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/izobrazevanje/podiplomski-doktorski-studij (Depth 1)
[INFO] [Worker-1] 2025-07-30 13:02:08 - [Thread Worker-1] Crawling (Depth 1): https://www.famnit.upr.si/sl/izobrazevanje/podiplomski-doktorski-studij
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/mednarodno-sodelovanje/izmenjave/studenti (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/latest (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/calendar (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/latest:rss (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/od-maja-do-sredine-p (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/raziskovalni-obisk-z-2 (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/raziskovalni-obisk-z-1 (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/kako-invazivna-zlata (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/zagnali-smo-aktivnos (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/vabilo-na-javno-pred-58 (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/vabilo-na-javno-pred-57 (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/novice/javni-poziv-za-prido-12 (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si//sl/novice/razpis-univerze-na-p-3 (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/bodoci-studenti/informativni-dan (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/bodoci-studenti/prijava (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/konference/ (Depth 1)
[INFO] [Worker-3] 2025-07-30 13:02:08 - Found new URL: https://www.famnit.upr.si/sl/konference (Depth 1)
```

Figure 1: Parallel Debug Output 1

```
[INFO] [Worker-3] 2025-07-30 13:00:25 - Skipped (different domain): mailto:referat@famnit.upr.si
[INFO] [Worker-3] 2025-07-30 13:00:25 - Skipped (different domain): https://www.facebook.com/up.famnit
[INFO] [Worker-3] 2025-07-30 13:00:25 - Skipped (different domain): https://plus.google.com/103521388605605631386
[INFO] [Thread-0] 2025-07-30 13:00:25 - Page limit reached. Shutting down.
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://www.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://www.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Found new URL: https://www.famnit.upr.si/sl/zaposleni-in-sodelavci/tjasa.ogrizek/ (Depth 3)
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): mailto:tjasa.ogrizek@famnit.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://splet02.izum.si/cobiss/bibliography?langbib=slv&li=si&homelang=svn&code=56138
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://splet02.izum.si/cobiss/bibliography?langbib=eng&li=en&homelang=svn&code=56138
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://e.famnit.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://e.famnit.upr.si/course/view.php?id=2005
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://intranet.famnit.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://matematicni-izleti.famnit.upr.si/sl/
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://bioloski-veceri.famnit.upr.si/sl/
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://naravoslovni-izleti.famnit.upr.si/sl/
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://matura.famnit.upr.si/sl/
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://advent.famnit.upr.si/
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): mailto:info@famnit.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): mailto:referat@famnit.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://www.iam.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://amc.imfm.si/index.php/amc%29
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://www.facebook.com/up.famnit
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://adam-journal.eu/index.php/ADAM/index
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): http://www.upr.si
[INFO] [Worker-8] 2025-07-30 13:00:27 - Skipped (different domain): https://plus.google.com/103521388605605631386
[INFO] [main] 2025-07-30 13:00:27 - Crawl completed. Total pages: 1007. Time: 71092 ms (71.09 seconds)
[INFO] [main] 2025-07-30 13:00:27 - Saved crawl log to crawled_links_par.txt
```

Figure 2: Parallel Debug Output 2

```
[INFO] 2025-07-30 12:15:28 - Starting crawl with:
       - Max Depth: 3
       - Max Pages: 1000
       - Domain: famnit.upr.si
[INFO] 2025-07-30 12:15:28 - Crawling (Depth 0): https://famnit.upr.si
[INFO] 2025-07-30 12:15:29 - Crawling (Depth 1): https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:29 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:29 - Crawling (Depth 2): https://www.famnit.upr.si
[INFO] 2025-07-30 12:15:31 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:31 - Skipping already visited or invalid URL: https://www.famnit.upr.si
[INFO] 2025-07-30 12:15:31 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:31 - Crawling (Depth 3): https://www.famnit.upr.si/en
[INFO] 2025-07-30 12:15:31 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:31 - Crawling (Depth 3): https://www.famnit.upr.si/sl/izobrazevanje/dodiplomski-studij
[INFO] 2025-07-30 12:15:31 - Crawling (Depth 3): https://www.famnit.upr.si/sl/izobrazevanje/podiplomski-magistrski-studij
[INFO] 2025-07-30 12:15:32 - Crawling (Depth 3): https://www.famnit.upr.si/sl/izobrazevanje/podiplomski-doktorski-studij
[INFO] 2025-07-30 12:15:32 - Crawling (Depth 3): https://www.famnit.upr.si/sl/mednarodno-sodelovanje/izmenjave/studenti
[INFO] 2025-07-30 12:15:32 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/latest
[INFO] 2025-07-30 12:15:33 - Crawling (Depth 3): https://www.famnit.upr.si/sl/calendar
[INFO] 2025-07-30 12:15:33 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/latest:rss
[INFO] 2025-07-30 12:15:34 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/od-maja-do-sredine-p
[INFO] 2025-07-30 12:15:34 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/raziskovalni-obisk-z-2
[INFO] 2025-07-30 12:15:34 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/raziskovalni-obisk-z-1
[INFO] 2025-07-30 12:15:34 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/kako-invazivna-zlata
[INFO] 2025-07-30 12:15:34 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/zagnali-smo-aktivnos
[INFO] 2025-07-30 12:15:34 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/vabilo-na-javno-pred-58
[INFO] 2025-07-30 12:15:35 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/vabilo-na-javno-pred-57
[INFO] 2025-07-30 12:15:35 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/javni-poziv-za-prido-12
[INFO] 2025-07-30 12:15:35 - Crawling (Depth 3): https://www.famnit.upr.si/sl/novice/razpis-univerze-na-p-3
[INFO] 2025-07-30 12:15:35 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl/novice/zagnali-smo-aktivnos
[INFO] 2025-07-30 12:15:35 - Crawling (Depth 3): https://www.famnit.upr.si/sl/bodoci-studenti/informativni-dan
[INFO] 2025-07-30 12:15:35 - Crawling (Depth 3): https://www.famnit.upr.si/sl/bodoci-studenti/prijava
[INFO] 2025-07-30 12:15:36 - Crawling (Depth 3): https://www.famnit.upr.si/sl/konference
[INFO] 2025-07-30 12:15:36 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:36 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl
[INFO] 2025-07-30 12:15:36 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl/konference
[INFO] 2025-07-30 12:15:36 - Crawling (Depth 3): https://conferences.famnit.upr.si/event/33
[INFO] 2025-07-30 12:15:37 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl/konference
[INFO] 2025-07-30 12:15:37 - Crawling (Depth 3): https://www.famnit.upr.si/sl/sis
[INFO] 2025-07-30 12:15:37 - Crawling (Depth 3): https://www.famnit.upr.si/sl/urniki
[INFO] 2025-07-30 12:15:37 - Crawling (Depth 3): https://www.famnit.upr.si/sl/email
[INFO] 2025-07-30 12:15:38 - Crawling (Depth 3): https://e.famnit.upr.si
[INFO] 2025-07-30 12:15:39 - Crawling (Depth 3): https://e.famnit.upr.si/course/view.php
[ERROR] 2025-07-30 12:15:39 - Retry 1 for URL: https://e.famnit.upr.si/course/view.php
[ERROR] 2025-07-30 12:15:41 - Retry 2 for URL: https://e.famnit.upr.si/course/view.php
[ERROR] 2025-07-30 12:15:43 - Retry 3 for URL: https://e.famnit.upr.si/course/view.php
[ERROR] 2025-07-30 12:15:45 - Failed to fetch URL after retries: https://e.famnit.upr.si/course/view.php
[INFO] 2025-07-30 12:15:45 - Crawling (Depth 3): https://www.famnit.upr.si/sl/obvestila
[INFO] 2025-07-30 12:15:47 - Crawling (Depth 3): https://intranet.famnit.upr.si
[INFO] 2025-07-30 12:15:48 - Crawling (Depth 3): https://www.famnit.upr.si/sl/studenti
[INFO] 2025-07-30 12:15:48 - Crawling (Depth 3): https://www.famnit.upr.si/sl/bodoci-studenti
[INFO] 2025-07-30 12:15:48 - Crawling (Depth 3): https://www.famnit.upr.si/sl/zaposleni-in-sodelavci
[INFO] 2025-07-30 12:15:49 - Crawling (Depth 3): https://www.famnit.upr.si/sl/gostujoci-profesorji
[INFO] 2025-07-30 12:15:50 - Crawling (Depth 3): https://www.famnit.upr.si/sl/o-fakulteti
```

Figure 3: Sequential Output (Depth 3) - Part 1

```
[INFO] 2025-07-30 12:15:57 - Skipping already visited or invalid URL: https://www.famnit.upr.si/sl/raziskovanje/programi-in-projekti/InnoRenew
[INFO] 2025-07-30 12:15:57 - Crawl completed. Total pages crawled: 54
[INFO] 2025-07-30 12:15:57 - Time taken for crawl: 29117 ms (29 seconds)
```

Figure 4: Sequential Output (Depth 3) - Part 2

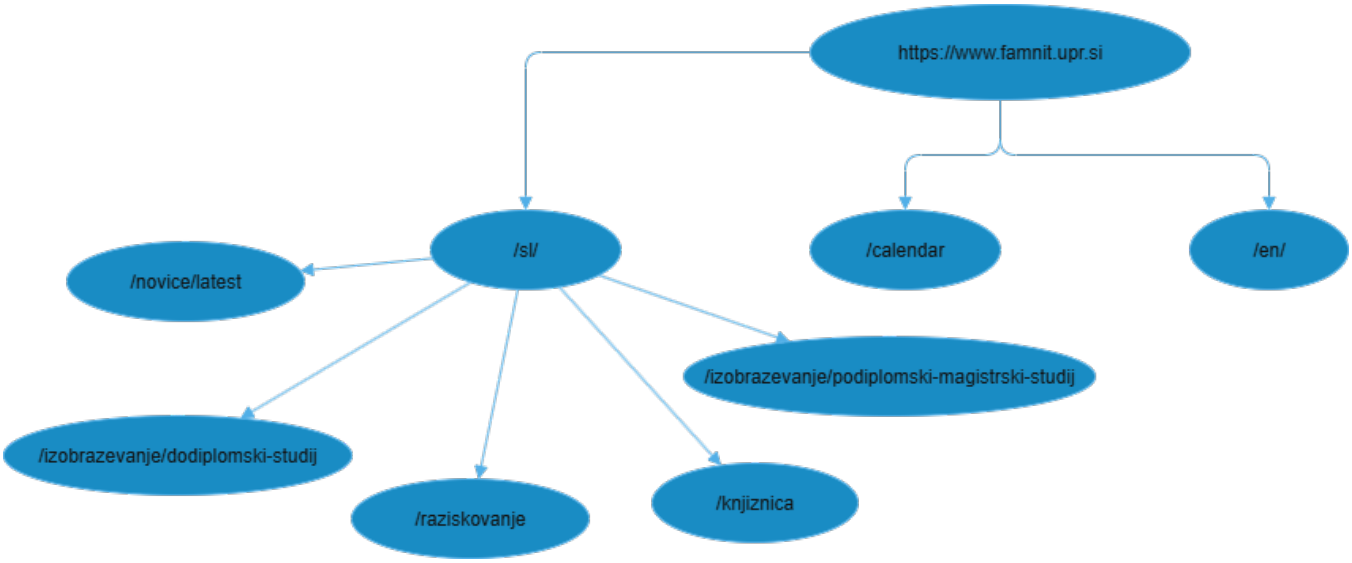

Figure 5: Crawl Structure Diagram