Focused Crawling for Structured Data

DSCI550

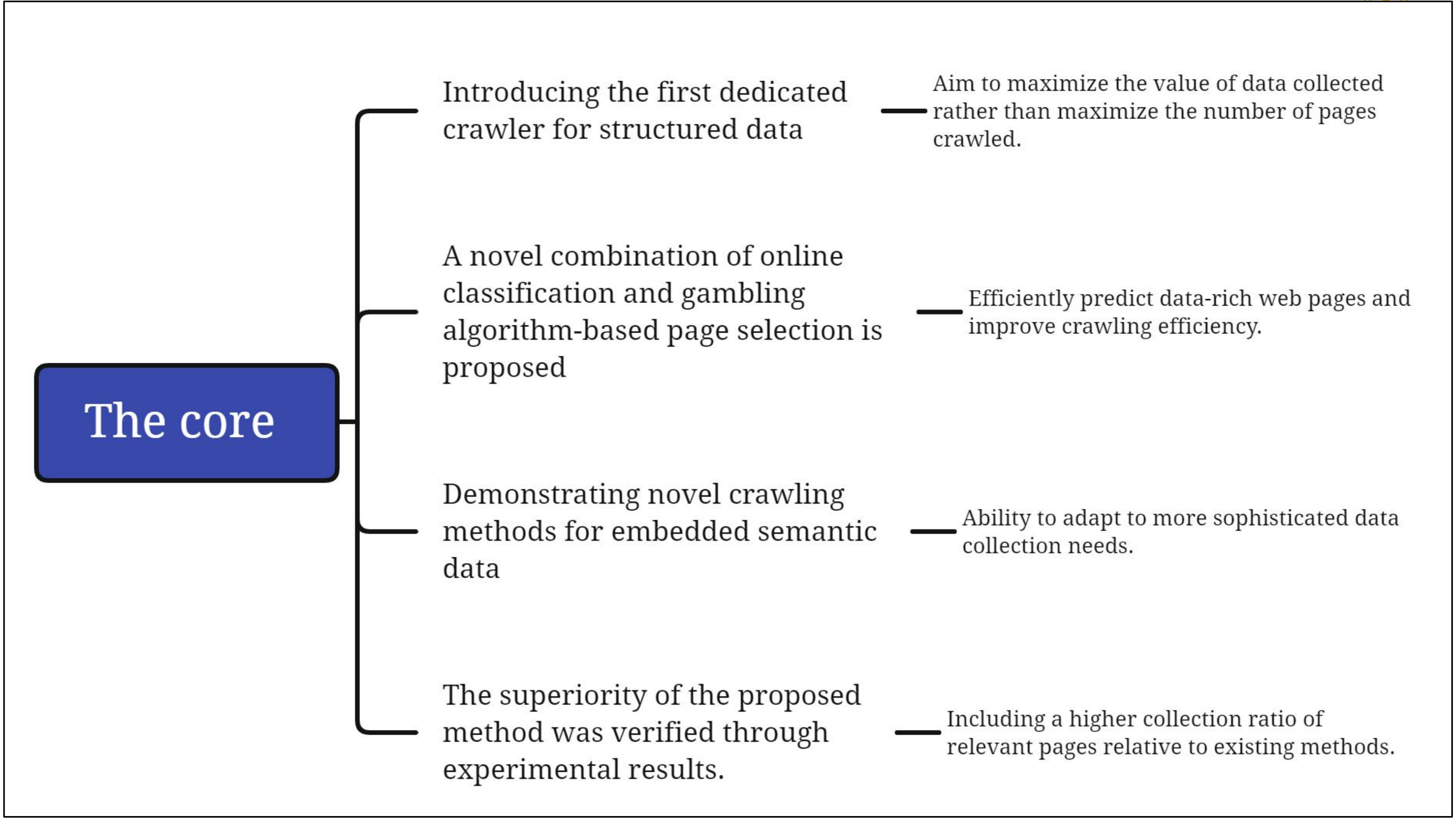
Presentation Paper

Yifan Yang

8386626967

yyang295@usc.edu

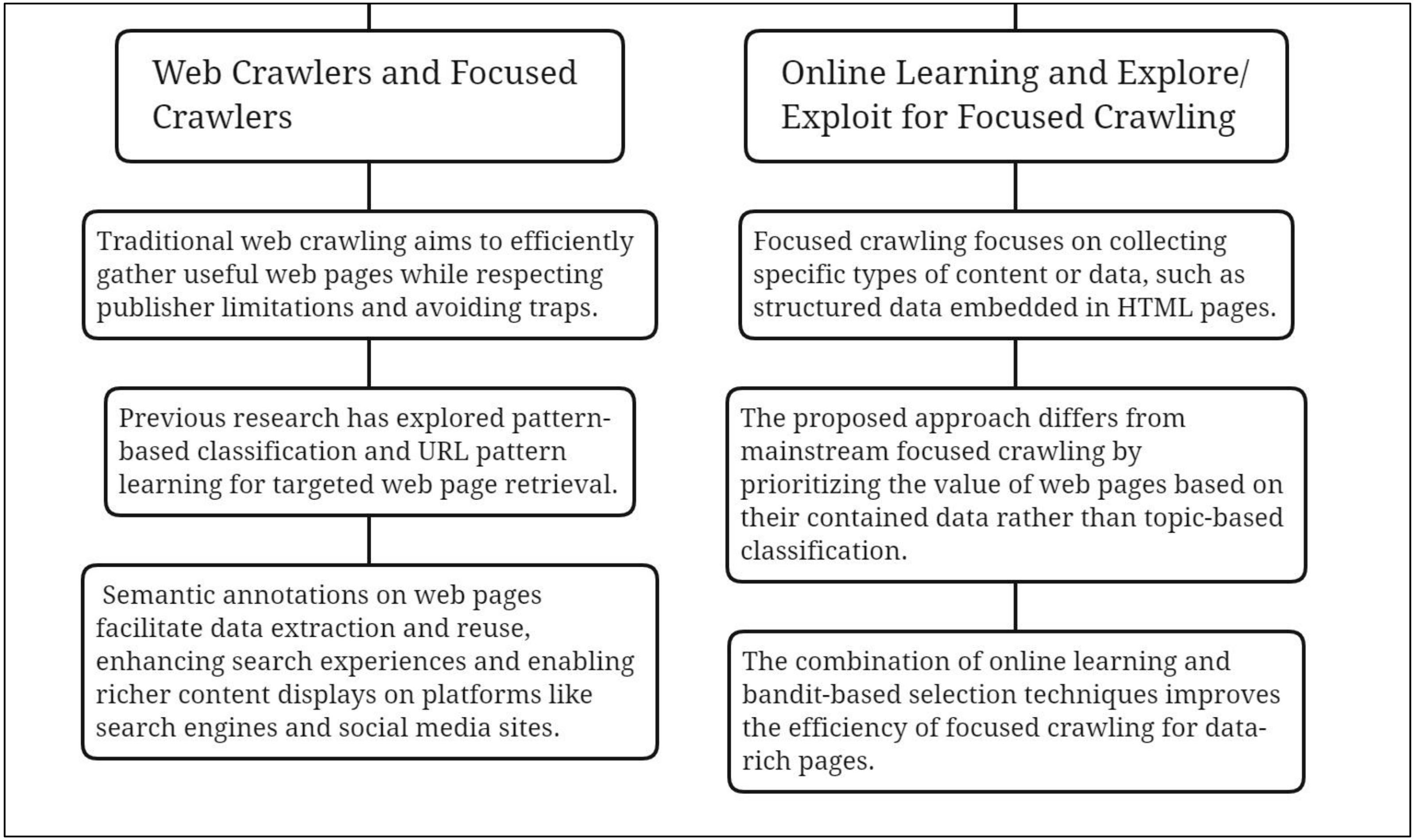
What I understand about this paper



The core

- Introducing the first dedicated crawler for structured data
  - Aim to maximize the value of data collected rather than maximize the number of pages crawled.
- A novel combination of online classification and gambling algorithm-based page selection is proposed
  - Efficiently predict data-rich web pages and improve crawling efficiency.
- Demonstrating novel crawling methods for embedded semantic data
  - Ability to adapt to more sophisticated data collection needs.
- The superiority of the proposed method was verified through experimental results.
  - Including a higher collection ratio of relevant pages relative to existing methods.

Discussion of related work and background



Web Crawlers and Focused Crawlers

Traditional web crawling aims to efficiently gather useful web pages while respecting publisher limitations and avoiding traps.

Previous research has explored pattern-based classification and URL pattern learning for targeted web page retrieval.

Semantic annotations on web pages facilitate data extraction and reuse, enhancing search experiences and enabling richer content displays on platforms like search engines and social media sites.

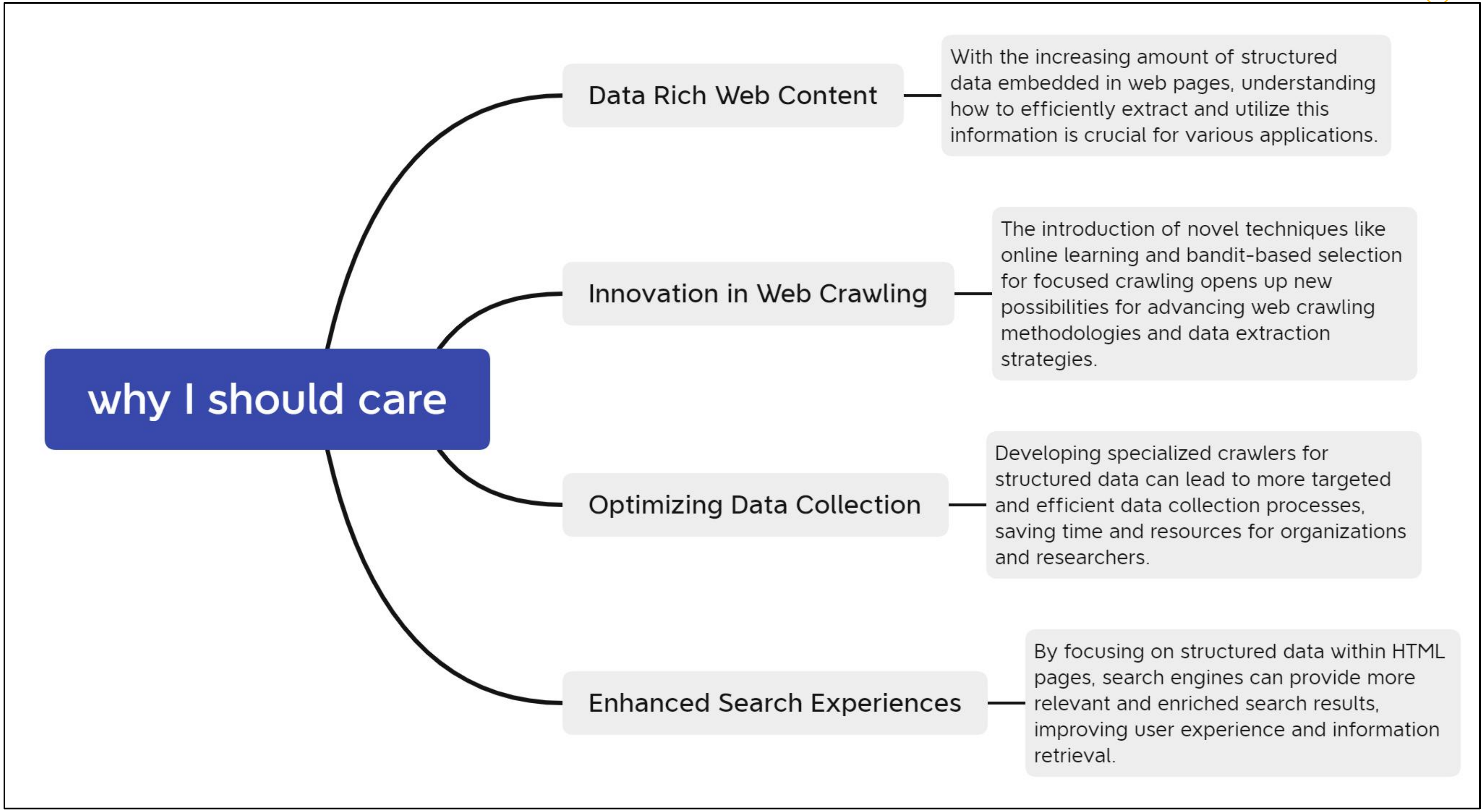
Online Learning and Explore/ Exploit for Focused Crawling

Focused crawling focuses on collecting specific types of content or data, such as structured data embedded in HTML pages.

The proposed approach differs from mainstream focused crawling by prioritizing the value of web pages based on their contained data rather than topic-based classification.

The combination of online learning and bandit-based selection techniques improves the efficiency of focused crawling for data-rich pages.

Discussion of why I should care about the topic



why I should care

Data Rich Web Content

With the increasing amount of structured data embedded in web pages, understanding how to efficiently extract and utilize this information is crucial for various applications.

Innovation in Web Crawling

The introduction of novel techniques like online learning and bandit-based selection for focused crawling opens up new possibilities for advancing web crawling methodologies and data extraction strategies.

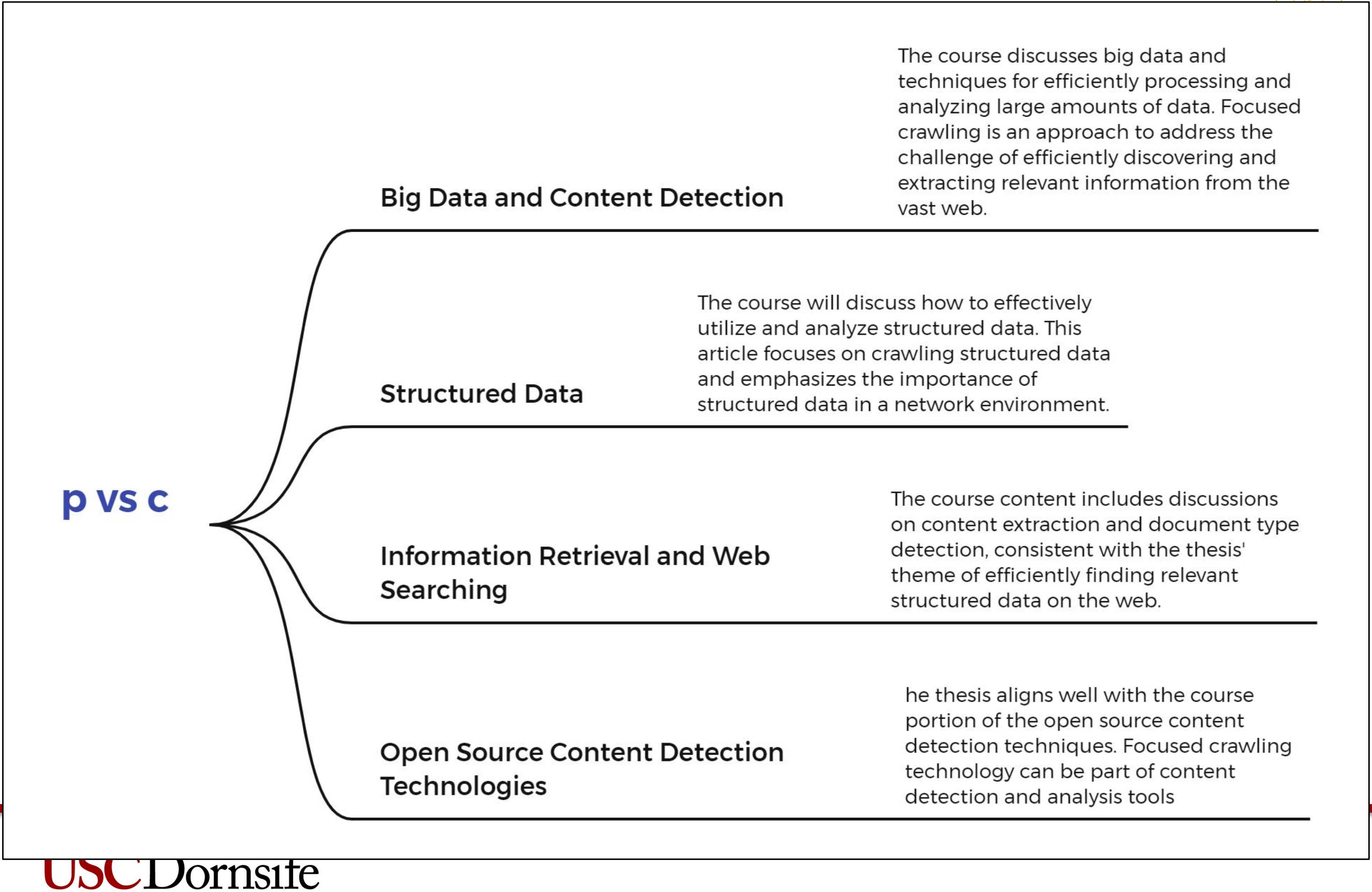
Optimizing Data Collection

Developing specialized crawlers for structured data can lead to more targeted and efficient data collection processes, saving time and resources for organizations and researchers.

Enhanced Search Experiences

By focusing on structured data within HTML pages, search engines can provide more relevant and enriched search results, improving user experience and information retrieval.

Relevant points between this paper and the class



p vs c

Big Data and Content Detection

The course discusses big data and techniques for efficiently processing and analyzing large amounts of data. Focused crawling is an approach to address the challenge of efficiently discovering and extracting relevant information from the vast web.

Structured Data

The course will discuss how to effectively utilize and analyze structured data. This article focuses on crawling structured data and emphasizes the importance of structured data in a network environment.

Information Retrieval and Web Searching

The course content includes discussions on content extraction and document type detection, consistent with the thesis' theme of efficiently finding relevant structured data on the web.

Open Source Content Detection Technologies

he thesis aligns well with the course portion of the open source content detection techniques. Focused crawling technology can be part of content detection and analysis tools

Summary and introduction of algorithms or technologies

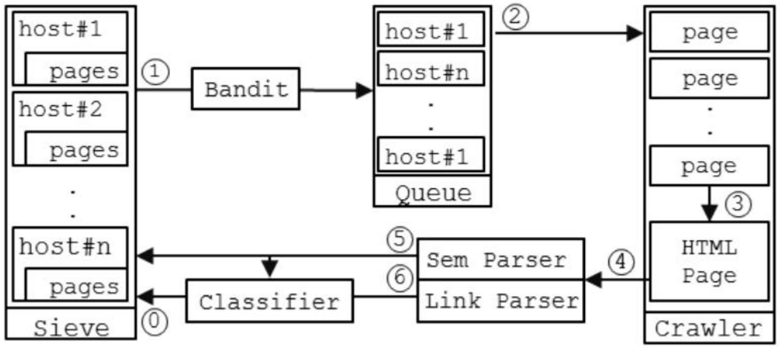
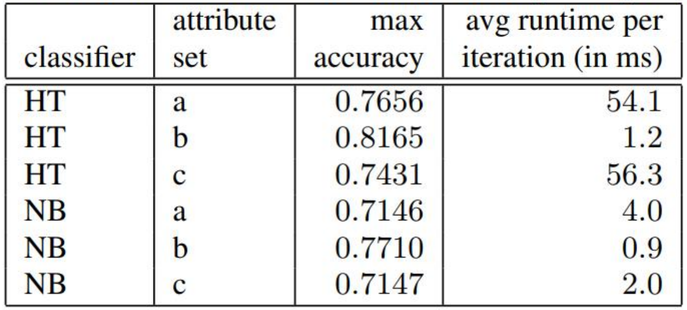
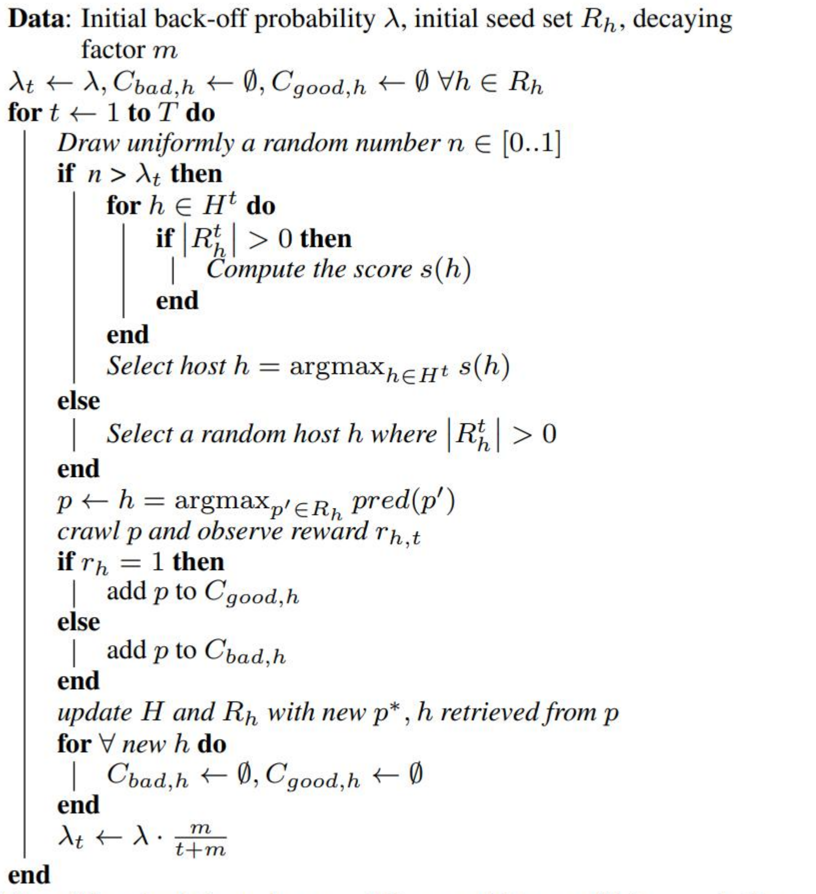


Figure 1: The architecture of Anthelion

Table 1: Results of feature and classification pre-experiments



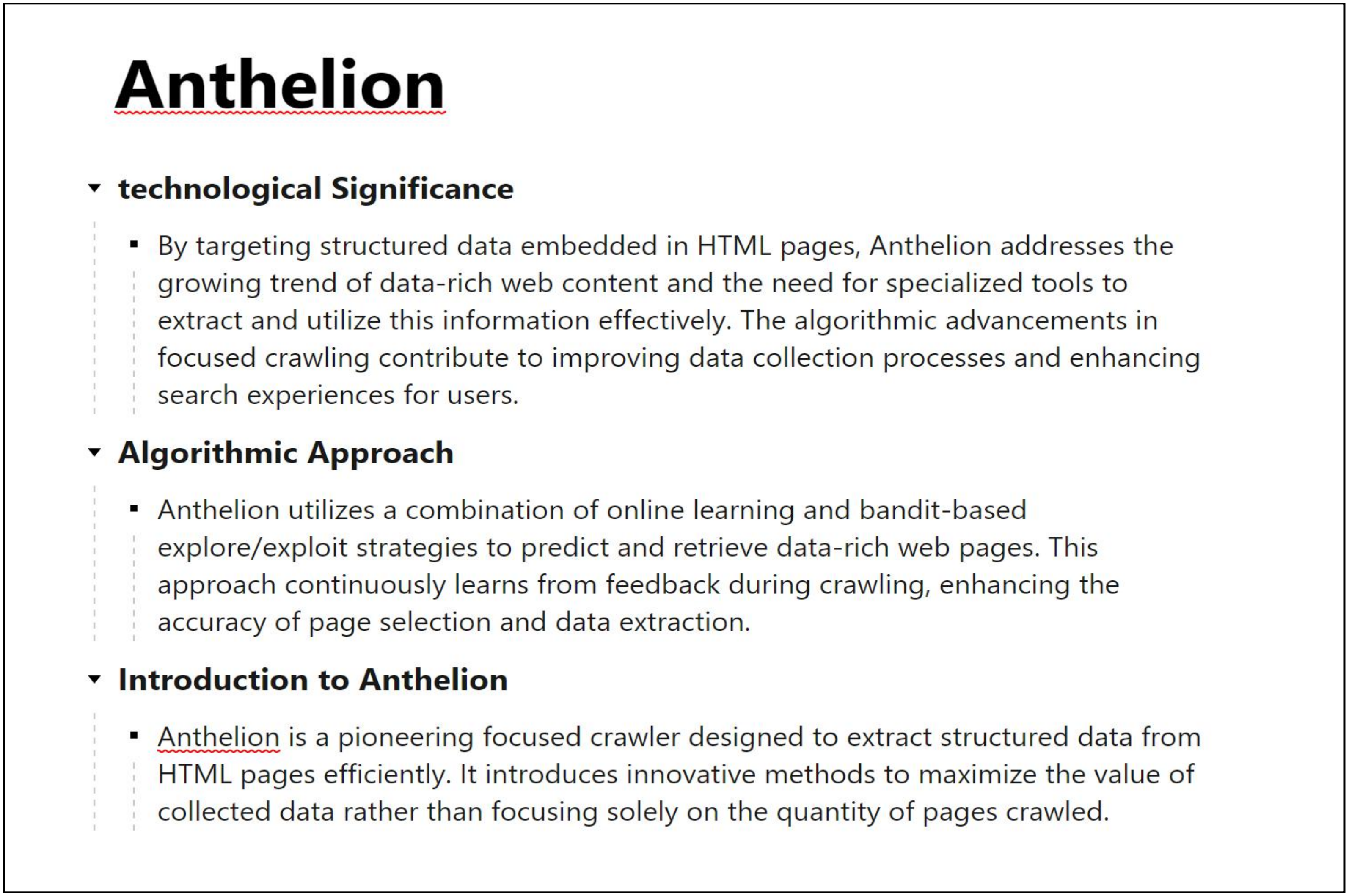
classifier	attribute set	max accuracy	avg runtime per iteration (in ms)
HT	a	0.7656	54.1
HT	b	0.8165	1.2
HT	c	0.7431	56.3
NB	a	0.7146	4.0
NB	b	0.7710	0.9
NB	c	0.7147	2.0



```
Data: Initial back-off probability $\lambda$, initial seed set $R_h$, decaying
      factor $m$
$\lambda_t \leftarrow \lambda, C_{bad,h} \leftarrow \emptyset, C_{good,h} \leftarrow \emptyset \; \forall h \in R_h$
for $t \leftarrow 1$ to $T$ do
    Draw uniformly a random number $n \in [0..1]$
    if $n > \lambda_t$ then
        for $h \in H^t$ do
            if $|R_h^t| > 0$ then
                Compute the score $s(h)$
            end
        end
        Select host $h = \text{argmax}_{h \in H^t} \; s(h)$
    else
        Select a random host $h$ where $|R_h^t| > 0$
    end
    $p \leftarrow h = \text{argmax}_{p' \in R_h} \; pred(p')$
    crawl $p$ and observe reward $r_{h,t}$
    if $r_h = 1$ then
        add $p$ to $C_{good,h}$
    else
        add $p$ to $C_{bad,h}$
    end
    update $H$ and $R_h$ with new $p^*$, $h$ retrieved from $p$
    for $\forall$ new $h$ do
        $C_{bad,h} \leftarrow \emptyset, C_{good,h} \leftarrow \emptyset$
    end
    $\lambda_t \leftarrow \lambda \cdot \frac{m}{t+m}$
end
```

Algorithm 1: Adapted general K-armed Bernoulli λ-greedy Bandit for focused crawling, with a linear decaying factor.

Summary and introduction of algorithms or technologies



Anthelion

technological Significance

- By targeting structured data embedded in HTML pages, Anthelion addresses the growing trend of data-rich web content and the need for specialized tools to extract and utilize this information effectively. The algorithmic advancements in focused crawling contribute to improving data collection processes and enhancing search experiences for users.

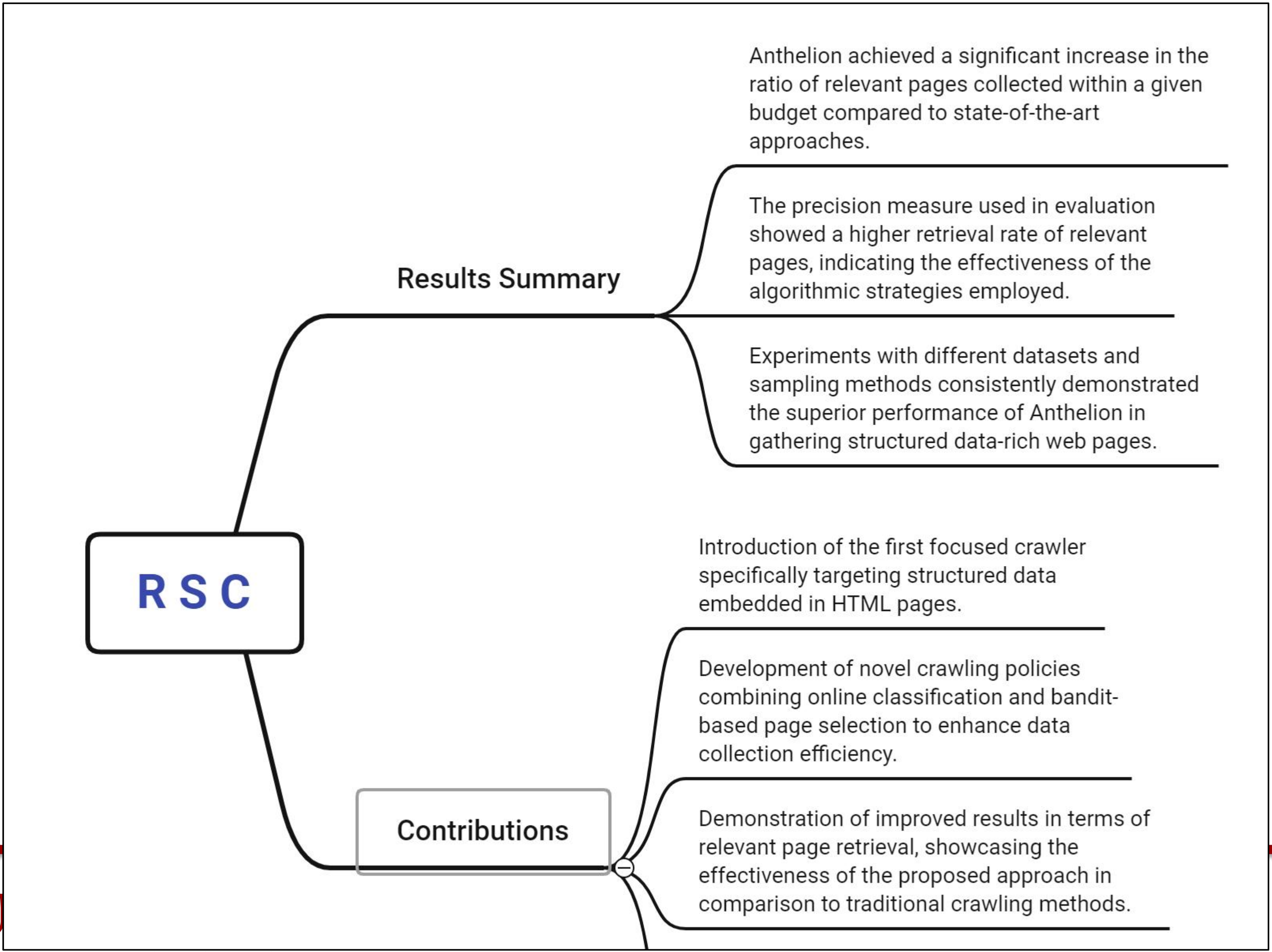
Algorithmic Approach

- Anthelion utilizes a combination of online learning and bandit-based explore/exploit strategies to predict and retrieve data-rich web pages. This approach continuously learns from feedback during crawling, enhancing the accuracy of page selection and data extraction.

Introduction to Anthelion

- Anthelion is a pioneering focused crawler designed to extract structured data from HTML pages efficiently. It introduces innovative methods to maximize the value of collected data rather than focusing solely on the quantity of pages crawled.

Results, summary and contributions of the paper



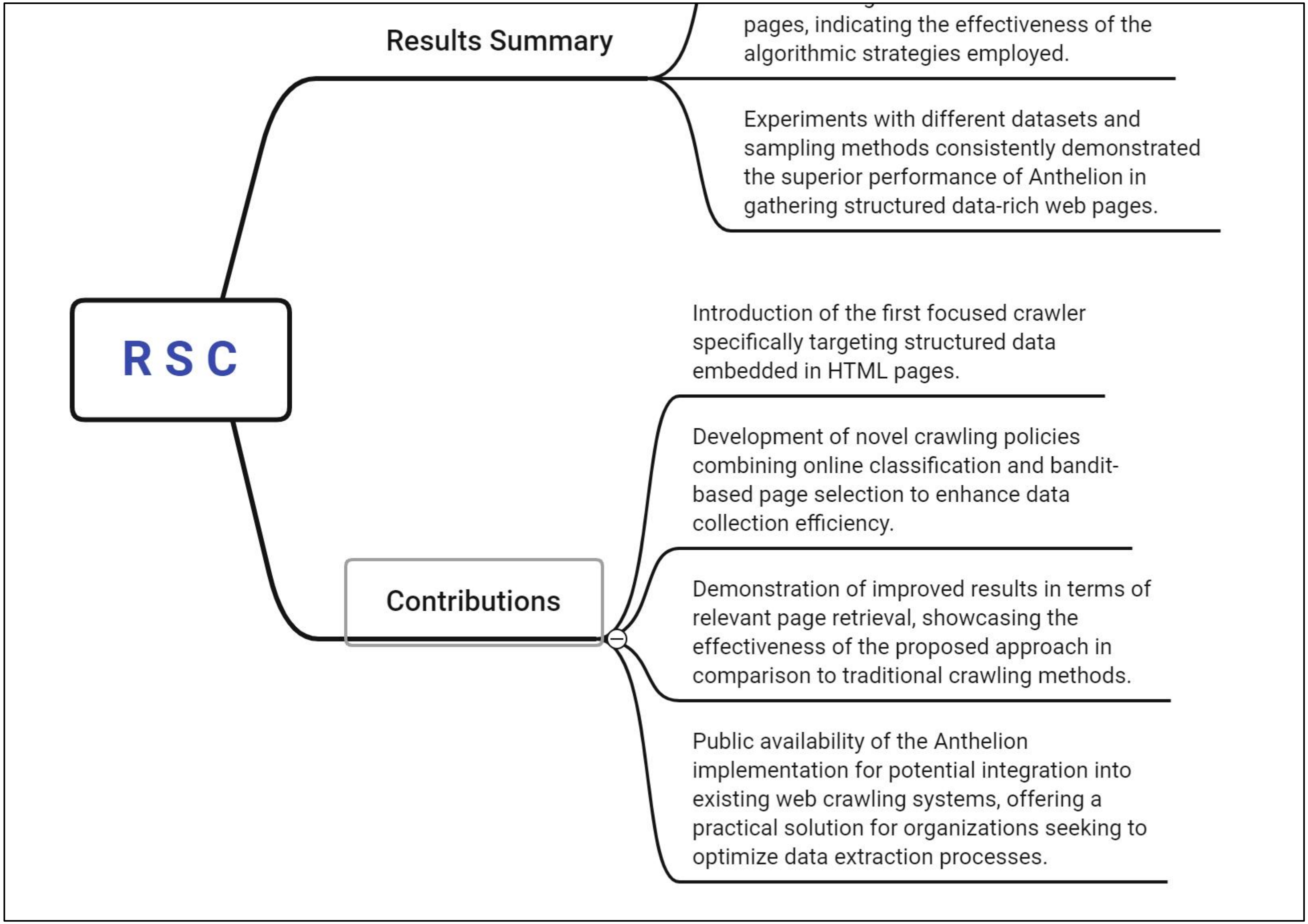
R S C

Results Summary

- Anthelion achieved a significant increase in the ratio of relevant pages collected within a given budget compared to state-of-the-art approaches.
- The precision measure used in evaluation showed a higher retrieval rate of relevant pages, indicating the effectiveness of the algorithmic strategies employed.
- Experiments with different datasets and sampling methods consistently demonstrated the superior performance of Anthelion in gathering structured data-rich web pages.

Contributions

- Introduction of the first focused crawler specifically targeting structured data embedded in HTML pages.
- Development of novel crawling policies combining online classification and bandit-based page selection to enhance data collection efficiency.
- Demonstration of improved results in terms of relevant page retrieval, showcasing the effectiveness of the proposed approach in comparison to traditional crawling methods.



RSC

Results Summary

pages, indicating the effectiveness of the algorithmic strategies employed.

Experiments with different datasets and sampling methods consistently demonstrated the superior performance of Anthelion in gathering structured data-rich web pages.

Contributions

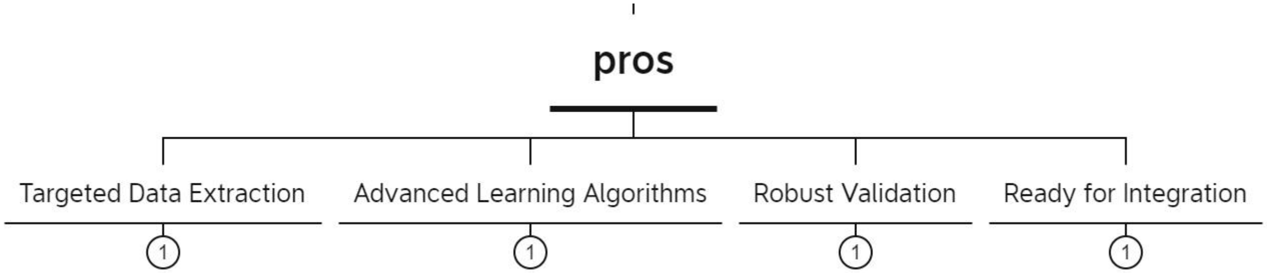
Introduction of the first focused crawler specifically targeting structured data embedded in HTML pages.

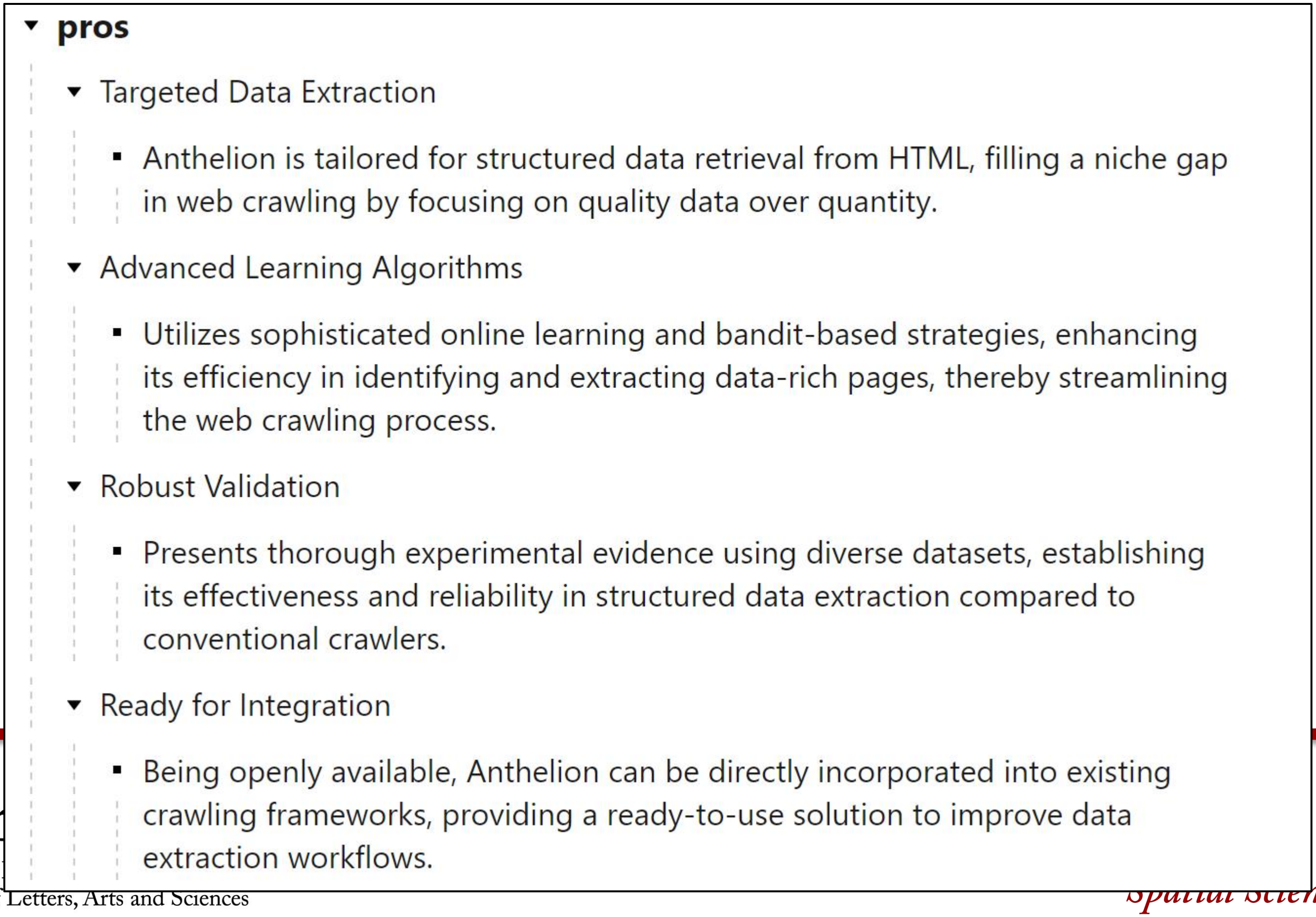
Development of novel crawling policies combining online classification and bandit-based page selection to enhance data collection efficiency.

Demonstration of improved results in terms of relevant page retrieval, showcasing the effectiveness of the proposed approach in comparison to traditional crawling methods.

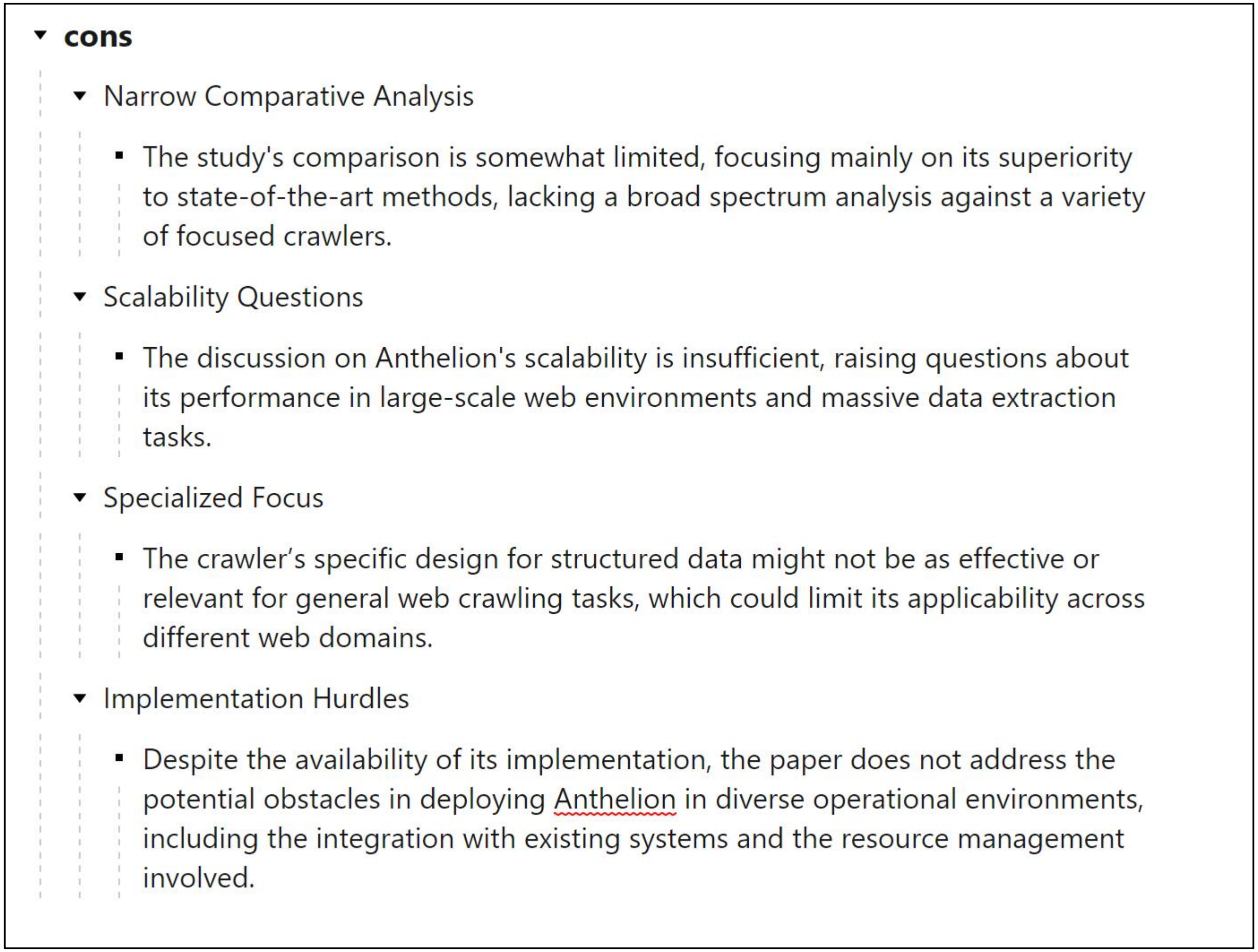
Public availability of the Anthelion implementation for potential integration into existing web crawling systems, offering a practical solution for organizations seeking to optimize data extraction processes.

Evaluate the pros and cons of the paper





- **pros**
  - Targeted Data Extraction
    - Anthelion is tailored for structured data retrieval from HTML, filling a niche gap in web crawling by focusing on quality data over quantity.
  - Advanced Learning Algorithms
    - Utilizes sophisticated online learning and bandit-based strategies, enhancing its efficiency in identifying and extracting data-rich pages, thereby streamlining the web crawling process.
  - Robust Validation
    - Presents thorough experimental evidence using diverse datasets, establishing its effectiveness and reliability in structured data extraction compared to conventional crawlers.
  - Ready for Integration
    - Being openly available, Anthelion can be directly incorporated into existing crawling frameworks, providing a ready-to-use solution to improve data extraction workflows.



cons

- Narrow Comparative Analysis
  - The study's comparison is somewhat limited, focusing mainly on its superiority to state-of-the-art methods, lacking a broad spectrum analysis against a variety of focused crawlers.
- Scalability Questions
  - The discussion on Anthelion's scalability is insufficient, raising questions about its performance in large-scale web environments and massive data extraction tasks.
- Specialized Focus
  - The crawler's specific design for structured data might not be as effective or relevant for general web crawling tasks, which could limit its applicability across different web domains.
- Implementation Hurdles
  - Despite the availability of its implementation, the paper does not address the potential obstacles in deploying Anthelion in diverse operational environments, including the integration with existing systems and the resource management involved.