IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁘ Vol. 9, Issue 3, March 2022

DOI: 10.17148/IARJSET.2022.9346

Focused Crawler: Uses and Challenges

Dr. Kompal

Govt. College, Panchkula

Abstract: A focused crawler, also known as a topical crawler or selective crawler, is a web crawler designed to index specific types of content or websites based on predefined topics or criteria. Unlike general-purpose web crawlers that aim to index as much content as possible across the web, focused crawlers are more targeted and efficient.

Keywords: Focused crawler or topical crawler or selective crawler, Natural Language Processing (NLP), Crawl Policy, Indexing.

1. INTRODUCTION

Focused crawlers typically start with a set of seed URLs that are known to belong to the target topic or domain. These seed URLs serve as the starting points for the crawling process. Focused crawlers use predefined relevance criteria to determine whether a webpage is relevant to the target topic. These criteria can include keywords, metadata (such as title, description, and tags), URL patterns, domain authority, and other indicators of topical relevance.

2. METHODOLOGY

Crawlers analyze the content of webpages to assess their relevance to the target topic. This analysis may involve techniques such as keyword extraction, topic modeling, Natural Language Processing (NLP), and machine learning algorithms to identify relevant content.

1. Crawlers employ various strategies to focus their crawling efforts on pages that are most likely to contain relevant content. These strategies may include:

- Breadth-first or depth-first traversal: Prioritizing pages based on their proximity to the seed URLs.
- Host-based crawling: Prioritizing pages from authoritative or trusted domains.
- Page quality assessment: Evaluating the quality of pages based on factors such as content freshness, popularity, and credibility.
- Topic drift detection: Monitoring changes in the topical relevance of crawled pages and adjusting crawling priorities accordingly.

2. Crawl Policy: Focused crawlers use crawl policies to govern the crawling process, including parameters such as crawl frequency, crawl depth, politeness (to avoid overloading servers), and revisit policies (to update previously crawled pages).

3. Some focused crawlers incorporate feedback mechanisms to iteratively refine their crawling strategies based on the relevance of previously crawled pages. This can involve user feedback, relevance feedback algorithms, or learning algorithms that adapt to changes in the web environment.

4. By employing various techniques, focused crawlers can efficiently discover and index relevant content while minimizing the resources required for crawling and maintaining topical relevance.

5. Indexing Relevant Content: Focused crawlers prioritize crawling and indexing that are most relevant to the target topic or domain. By focusing on relevant content, the crawler ensures that the search engine's index contains a higher proportion of pages that are likely to be useful to users searching for information on that topic.

6. Supporting Freshness and Diversity: Focused crawlers can help search engines maintain a fresh and diverse index by continuously discovering and indexing new content relevant to the target topic. This freshness and diversity of content can positively impact search engine rankings, as search engines strive to provide users with up-to-date and varied search results.

DOI: 10.17148/IARJSET.2022.9346

7. Reducing Noise and Spam: Focused crawlers are designed to avoid crawling irrelevant or low-quality content, such as spam, duplicate pages, or thin content. By filtering out noise and spam, the crawler helps ensure that the search engine's index contains higher-quality content, which can improve the overall quality of search results and enhance user satisfaction.

8. Optimizing Crawling Efficiency: Focused crawlers optimize crawling efficiency by prioritizing resources on crawling and indexing pages that are most likely to contribute to improved search engine rankings. By efficiently discovering and indexing relevant content, the crawler enables search engines to allocate resources more effectively and focus on delivering high-quality search results to users.

3. CHALLENGES

1. **Relevance and Precision**: Ensuring that the crawled content is highly relevant to the specified topic or domain is a fundamental challenge. Focused crawlers must employ sophisticated algorithms to prioritize URLs and content that are likely to be relevant, while filtering out irrelevant or redundant information.

2. **Scalability and Efficiency**: Focused crawling often involves traversing large portions of the web to locate relevant content. Maintaining scalability and efficiency in dynamic environment requires optimization of crawling strategies, resource management, and distributed crawling techniques.

3. **Dynamic Content**: The web is constantly evolving, with new pages being created, existing pages being updated, and obsolete content being removed. Focused crawlers must adapt to these changes by revisiting previously crawled pages, detecting content updates, and identifying new relevant sources.

4. **Handling Multilingual and Multimodal Content**: The web contains content in multiple languages and formats, including text, images, videos, and multimedia. Focused crawlers need to be capable of processing and understanding diverse types of content to ensure comprehensive coverage of the targeted topics or domains.

5. **Spam and Low-Quality Content**: The presence of malicious content on the web poses a significant challenge for focused crawlers. Such content can pollute the crawl results, reduce the quality of the collected data. Effective spam detection and filtering mechanisms are essential to mitigate this challenge.

6. **Crawling Depth and Coverage**: Determining the optimal crawling depth (i.e., how many levels deep to crawl from seed URLs) and coverage (i.e., the proportion of relevant content retrieved) poses a challenge in focused crawling. Large coverage is essential to maximize the retrieval of relevant information while minimizing redundant or irrelevant content.

7. **Adaptability to User Preferences and Feedback**: Incorporating user preferences and feedback into the crawling process can enhance the relevance and effectiveness of focused crawlers.

REFERENCES

[1] Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd.Annual Workshop on the Weblogging Ecosystem, 2005
[2] Qingyang Xu , Wanli Zuo, "First-order Focused Crawling", International World Wide Web Conferences 2007
[3] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios, "Improving the Performance of Focused WebCrawlers", Data & Knowledge Engineering, Vol: 68, No: 10, pp: 1001-1013, October 2009
[4] Dilip Kumar Sharma , A.K. Sharma," A Comparative Analysis of Web Page Ranking Algorithms", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670- 2676
[5] Swati Mali, B.B.Meshram, "Focused Web Crawler with Page Change Detection Policy" in 2nd International Conference and workshop on Emerging Trends in Technology (ICWET) Proceedings published by International Journal of Computer Applications® (IJCA), 2011
[6] Chain Singh, Ashish Kr. Luhach, Amitesh Kumar, "Improving Focused Crawling with Genetic Algorithms", International Journal of Computer Applications , March 2013.