IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Survey Significance of AI for ensuring precise Web Crawling

Mr. Lokesh M[1]
Assistant professor
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Mr. Murgan S[2]
Assistant professor
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Jayanth B R[3]
(5th Sem , 3rd year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Akanksha Singh[4]
(5th Sem, 3rd year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

K.Rekha[5]
(5th Sem, 3rd year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Kailash Kumar[6]
(5th Sem, 3rd year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

ABSTRACT:

This compilation analyzes the application of artificial intelligence in web indexing, covering diverse sectors including crawler classifications, profound learning models, insightful frameworks, and moral contemplations. This investigation features key discoveries identifying with the viability of unique crawler sorts, the utilization of factual models for anticipating out connections, and the advancement of occasion focused looking frameworks. The audit additionally stresses the significance of clearness, pertinent data, achievability, quantifiability, and good conduct consistency in web scraping ventures. All in all, this wide ranging assessment shows the huge potential for AI to upgrade web looking and information investigation for an assortment of uses. The report incorporates both long, intricate passages and short, direct sentences to catch various points of view on this creating field.

Keywords: Crawler, CAIMANS, NLP

INTRODUCTION

[1] The intricate web of pages spanning the internet hold untold information within their depths, but extracting what is pertinent requires strategies to efficiently traverse this vast network. This investigation aims to examine focused crawlers, inference-based crawlers, incremental crawlers, parallel crawlers, and distributed crawlers; evaluating

each approach's ability to scour sites and uncover carefully what is germane. [2] Convolutional neural networks show promise for part-of-speech tagging in Marathi, a language with intricate grammar. Yet how do their capabilities compare when analyzed alongside more conventional machine learning algorithms? This study looks to contrast effectiveness by assessing the accuracy of a suggested bi-directional long short-term memory model. [3] To grapple with the challenges of harvesting digital artifacts from the expanse of the web and connecting them to internal reservoirs of big data and accumulated knowledge, the intelligent CAIMANS system was developed. Ever seeking to enhance its capabilities, researchers explore avenues such as large-scale data techniques, machine learning methods, and natural language processing.

Extend the system's applicability beyond e-procurement to other domains as real-world use cases and comparison studies validate CAIMANS' efficacy across various application areas. Validate the CAIMANS system through real-world use cases and a comparison study with other systems currently used in several application areas to demonstrate its efficacy and usability. [3] [4] Clearly define focused yet flexible research objectives that meaningfully contribute to overall goals while aligning with technical capabilities, ethically available data sources, and quantifiable project standards assessing success. Ensure research objectives contribute value by establishing relevance with overarching aims through flexible yet focused definitions and quantifiable success measures considering technical, data, and ethical feasibilities.

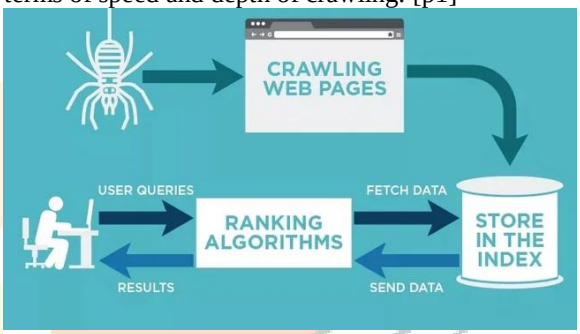
Ethical Compliance: Adhere to ethical standards and legal regulations in data collection to maintain integrity and credibility in the research process.[4] [5] The primary goal of this research is to analyze the effective predictors of new outlines in focused web crawling to enhance web analytics for small and medium enterprises. The purpose of this project is to use statistical models and feature design to forecast how many new outlines will be posted on a website.[5]

[6] The research aims to develop an integrated event-focused crawling system to efficiently gather web data related to key events.[6] This research aims to develop an intelligent system to support organizations in crawling artifacts from the web, semantically matching them against internal Big Data sources, and enabling better strategic decision-making, especially in the e-procurement domain. The system extends the k-means algorithm for web crawling and incorporates a semantic module for data analysis.[7]. [8] This study investigated emotional attitudes toward procrastination and explored the predictive role of these sentiments on procrastination predisposition. Specific objectives included analyzing sentiments toward procrastination, examining demographic differences in sentiment evaluation, exploring potential interactions among emotional attitudes, and predicting procrastination predisposition based on sentiments.[8] [9] This research aims to develop a context-aware recommender system for restaurant recommendations using web crawling techniques and deep learning models. The specific objectives include extracting contextual features from user feedback, building an information representation model using word embeddings, and implementing a deep recurrent neural network in place of recommendation predictions.[9]

LITERATURE REVIEW

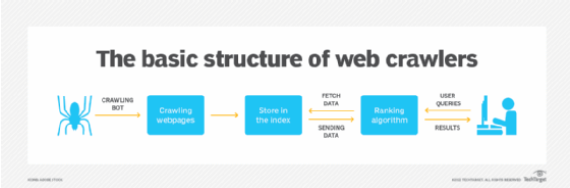
[p1] The researchers employed a survey approach to gather information about the features, pros, and cons of various crawler architectures. They analyzed the working methodologies of different crawlers and compared their performance Regarding efficiency, speed, and relevancy in accessing web content.

Focused crawlers are efficient for obtaining results on specific topics and can be scaled to access hidden web content. Incremental crawlers excel in maintaining an updated database but may lack the ability to crawl deep web content. Inference–based crawlers produce better search results by analyzing previously traversed links; however, they may have limitations in terms of speed and depth of crawling. [p1]



[p2] Creation of a POS–tagged corpus with 1500 Marathi sentences using 32 tags. Implementation of the Bi–LSTM model for POS tagging. Evaluation of model performance using recall, accuracy, and F1 score metrics. The Bi–LSTM model attained a precision of 85% for POS tagging in Marathi text. Conventional machine learning methods were surpassed by deep learning for POS tagging. The model showed promising results in accurately identifying and labeling punctuation symbols.[p2] [p3] The methodology involves investigating and implementing Natural language processing, huge data, and machine learning techniques to enhance the CAIMANS system. This includes developing an advanced web crawler with an extended K-means algorithm, creating a semantic module for content analysis, extending the system to other domains, and validating its effectiveness through real-world use cases and comparative analysis. The creation of an intelligent system, CAIMANS, to support organizations in crawling web artifacts and matching them with internal data sourcesEffective application of cutting-edge methods to improve the system's capabilities, such as semantic analysis and expanded K-means clustering. system expansion into new application areas outside of e-procurement. Validation of the system's effectiveness and usability through real-size use cases and a comparison with current systems in different application domains [p3] [p4] Methodology in web scraping research involves outlining the systematic approach and techniques used to collect, extract, and analyze data from online sources. This includes defining the scope of the study, selecting appropriate web scraping tools and technologies, identifying target websites, and determining data extraction methods. The

methodology also encompasses considerations for data preprocessing, storage, and analysis, additionally ethical and legal aspects of web scraping. Researchers may employ programing languages such as Python, use web scraping libraries such as BeautifulSoup or Scrapy, and implement strategies for handling dynamic content and anti-scraping mechanisms. A well-defined methodology ensures the reliability, validity, and efficiency of the web scraping process. Web scraping is a valuable tool in order to retrieve relevant info from unstructured sources, with applications in business intelligence, data science, big data, and cybersecurity. Python is a popular language for web scraping because of its simplicity, readability, and strong community support. Ethical and legal aspects are important in web scraping, and compliance with website terms of use, copyright laws, and ethical practices is crucial. In addition to ongoing legal and ethical debates, online scraping may see a rise in the use of AI and machine learning in the future. [p4]



[p5] This study involves a systematic analysis of web change prediction, statistical modeling, and feature design. A new dataset of ten weekly crawls is used for empirical analysis and prediction modeling.

Feature design is categorized into static page features, dynamic page features, static network features, and dynamic network features. Statistical models were trained to predict the link change rate, presence of new outlooks, and number of new outlines on a page.

The NG Boost method is employed to predict the average number of new outlines according to a Poisson distribution assumption. Features related to page content, history of changes, incoming hyperlinks, and TrustRank are identified as important predictors. This research offers perceptions into the predictability of new outlines in focused crawling.

The feature design taxonomy and statistical modeling techniques are discussed in detail.

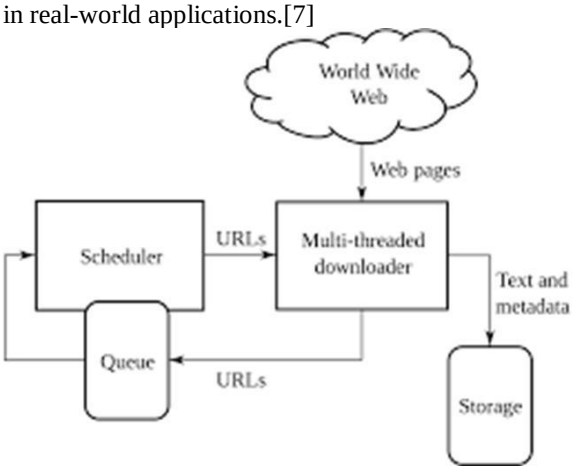
The practical relevance of the Poisson distribution assumption for predicting new outlines is explored.

[**6**] This study proposes a keyword weight optimization method utilizing the stochastic gradient descent (SGD) to enhance keyword sets in event-focused web crawling. The methodology uses SVM classifiers and term frequency-based feature extraction to improve the relevancy of unvisited URLs.

1. The proposed method effectively optimizes keyword weights for focused web crawling during important events.
2. The application of SGD and SVM classifiers enhances the relevancy determination of unvisited URLs.
3. This study demonstrates the importance of keyword optimization in improving the standard of keyword collections for web crawling.

[7] The system uses advanced technologies such as natural language processing, data clustering, and semantic analysis to extract relevant information from crawled artifacts. It employs techniques such as query expansion and faceted search to enhance the retrieval of data. The system's effectiveness is evaluated through real-size use cases and comparative analyses with existing systems.

1. The extended k-means algorithm improves the segmentation of the web search results, aiding in better data analysis.
2. The semantic module enhances information retrieval by analyzing and extracting pertinent data from crawled web pages.
3. The system's validation with industrial stakeholders in the e-procurement domain demonstrates its effectiveness and usability in real-world applications.[7]



[8] The research used large-scale web scraping analysis and Using natural language processing for sentiment analysis (NLP). Data were collected from 488,601 registered users who expressed concern regarding procrastination. Structural Equation Modeling (SEM) was employed to predict procrastination predisposition over a decade. Procrastination was predominantly viewed negatively. Significant interactions were observed between age-gender and gender-education, with educated-less males and young adult males expressing more negative emotions toward procrastination. A two-class model of latent growth was used to identify adaptive and maladaptive procrastination patterns. Sentiments toward procrastination are predictive of procrastination predisposition over time. [8]

[9] The methodology involves web crawling to extract user feedback dataword embedding methods for characteristic extraction, and a deep recurrent neural network designed to offline knowledge building and recommendation predictions. This study also incorporates sentiment clustering as well as deep learning techniques to enhance recommendation accuracy. This study presents a novel approach for making context-aware recommendations in the restaurant domain by leveraging deep learning and sentiment analysis techniques. The Deep Recurrent Neural Network Optimization model significantly improves recommendation accuracy by considering contextual features extracted from user feedback data. [9]

RELATED WORK

[1] Strengths: This study provides a detailed analysis of different types of crawlers and their functionalities. It offers insights into the drawbacks and restrictions on existing crawlers, paving the way for future improvements in search engine development .

Weaknesses/Limitations: The study may lack in-depth empirical data or case studies to support the findings. The survey approach may have limitations in capturing the real-time performance metrics of crawlers in practical scenarios.[1] [2] Strengths: Deep learning models are used to increase POS tagging accuracy. Creation of a POS–tagged corpus specific to Marathi

Detailed evaluation metrics, such as precision, F1 and recall, score, are provided.

Weaknesses: Slightly more talk on the difficulties of applying deep learning models to Marathi text. Insufficient POS tagging compared with cutting-edge deep learning algorithms.The

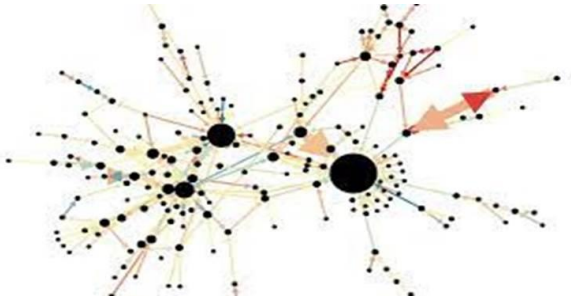
limited sample size of the POS–tagged corpus may affect generalizability. [2]

[3] Strength: CAIMANS demonstrates strength in leveraging cutting-edge technologies for efficient web artifact crawling and analysis. Validation through real-world use cases underscores its usability and effectiveness in supporting organizational decision-making. The system's scalability across various application domains showcases its adaptability and value in different industry settings. Overall, CAIMANS distinguishes itself with its sophisticated talents, validated performance, and potential to enhance decision-making processes within organizations across different sectors. Weakness**:** Despite its strengths, CAIMANS may have limitations such as potential challenges in handling complex and dynamic web environments, which could affect The precision of crawled data. The system's reliance on predefined rules and algorithms may limit its adaptability to evolving web content structures. Additionally, The system's efficacy could be influenced by the quality and relevance of the initial query inputs. Furthermore, the scalability of CAIMANS to extremely large datasets and diverse web sources may pose challenges in maintaining optimal performance and efficiency, potentially impacting its usability in certain contexts. [3] [4] Strength: Web scraping's strength is its capacity to effectively extract important data from unstructured sources, facilitating well-informed decision-making and offering insights into a variety of industries, including cybersecurity, data science, and business intelligence. Python's simplicity and strong community support make it a popular language for web scraping, facilitating the development of effective scraping tools and techniques. Weakness**:** The weaknesses and limitations of web scraping include legal and ethical considerations, potential copyright infringement, and the need for continuous monitoring to ensure compliance with website terms of use, copyright laws, and ethical practices. In addition, Python's asynchronous processing capabilities may be limited, and it may face challenges in handling complex web scraping tasks at scale. [4]

[5] Thorough analysis of web change prediction, systematic feature design taxonomy, and the application of a new dataset for empirical analysis. Limited discussion of the generalizability of the findings, potential biases in the dataset, and external validity of the models. This study focuses on a specific type of change (addition of new outlinks) and might not record every detail of alterations to a Web page. [5] [6] This study introduces a novel approach to keyword weight optimization in event-focused web crawling . This study incorporates machine learning techniques to increase the significance of crawled web data. The methodology provides a systematic way to improve keyword collections for improved data extraction and search engine indexing. This study may lack a detailed comparison with other state-of-the-art keyword optimization methods in web crawling. 2. This study might benefit from a larger-scale, more thorough assessment of the suggested methodology. 3. The limitations of this study may include the specific focus on event-driven web crawling, potentially limiting its applicability to general web crawling scenarios.[6]

[7]Strengths: Novel approach combining data crawling, semantic analysis, and Big Data technologies. Real–world validation with industrial stakeholders. Extensive evaluation and comparison with existing systems.

Weaknesses



1. Complexity of the CMIS Query Language for Users with Limited Computer Science Background

2. Need for continuous updates and enhancements to incorporate the latest techniques in data analysis.

3. Limited discussion on scalability and potential challenges in deploying the system across different domains.[7] [8] Strengths: use of large-scale data for comprehensive analysis. Integration of sentiment analysis and SEM for predictive modeling. Clear articulation of the research objectives and hypotheses.

Weakness: Reliance on web-scraped data may introduce bias. Lack of qualitative data to supplement sentiment analysis. Limited generalizability due to specific user demographics.[8] [9] Strengths: Innovative integration of web crawling, sentiment analysis, and deep learning for context-aware recommendations. Clear explanation of the methodology and model development process. Demonstrated improvement in recommendation accuracy using the optimized deep recurrent neural network. Weaknesses: Limited discussion on the scalability and real-time applicability of the proposed model. Lack of detailed comparison with state-of-the-art recommendation systems in other domains.

Potential challenges in generalizing the findings to different types of recommendation tasks beyond restaurant recommendations.[9]

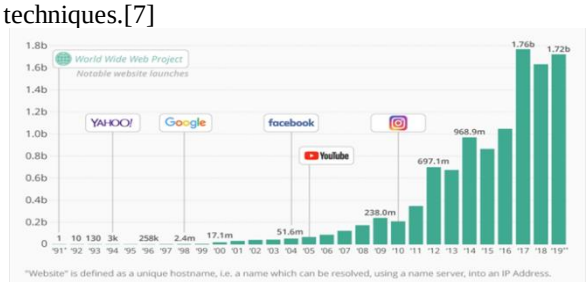
PROPOSED WORK

Examination of Crawling Methodologies: This section delves into the intricate world of web crawlers, dissecting different approaches like focused, inferencebased, incremental, parallel, and distributed crawlers. You'll embark on a journey through existing literature, surveys, and comparative studies to meticulously evaluate their strengths, weaknesses, and unique characteristics. This in-depth analysis will give you, at age 19, a comprehensive understanding of how each methodology navigates the vast web landscape and retrieves relevant information. Deep Learning for Marathi Text POS Tagging: 6 Get ready to see the incredible power of deep learning. unfold as you leverage Bi-LSTM models to tackle the intriguing task of Part-of-Speech (POS) tagging in Marathi text. This section meticulously compares the execution of Bi-LSTM against traditional machine learning algorithms, using meticulous assessment metrics like as 23 recall, precision, and F1 score to paint a clear picture of accuracy and efficiency. You'll gain insightful observations 1 on deep learning's potential for natural language 20 processing tasks in lesser-resourced languages. Development of Intelligent Systems for Web Crawling and Data Matching: Join the quest to conquer the challenges of web crawling and data matching across diverse application domains! This section examines the fascinating CAIMANS system, which demonstrates the potency of combining natural language processing, huge data, and machine learning. (NLP). Witness firsthand how these technologies elevate system capabilities through realworld use

cases and insightful comparisons with existing systems. CAIMANS promises to occupy game- shifter in the data integration field and analysis. Web Scraping and Sentiment Analysis: Brace yourself for an exciting exploration of methodologies that seamlessly blend web scraping and sentiment analysis to unlock the hidden treasures of online information. This section dissects the skill of extracting meaningful insights from diverse online sources, using Structural Equation Modeling (SEM) to unravel the fascinating connection between sentiments and procrastination predisposition. You'll discover how demographic differences affect sentiment evaluation, revealing The subtleties of human conduct in the digital age. [1] This study discusses the evolving challenges faced by crawlers due to the increasing complexity and ambiguity of web content. This highlights the need for developing smart crawlers capable of traversing diverse webpages and producing accurate search results.[1] [2] Comparison with machine learning techniques highlighted the superiority deep learning for POS tagging. The application of LSTM and GRU to address the vanishing gradient problem in RNNs is discussed. The importance of labeled data and neural network architectures in models for deep learning have been emphasized.[2] [3] This research paper introduces CAIMANS, an intelligent system that makes use of cutting-edge technology for web artifact crawling and analysis. Validation through real-world use cases demonstrates its usability and possible influence on decision-making processes. The system's scalability across different domains highlights its value in supporting organizations in strategic decision-making. [3] [4]Web scraping is a valuable tool for extracting data from unstructured sources, which is beneficial in fields such as business intelligence and cybersecurity. Python's simplicity and strong community support make it a popular language for web scraping. Legal and ethical considerations are crucial in web scraping, with compliance, copyright, and ethical practices being key factors.[4] [5] This work sheds light on how predictable fresh outlines in focussed crawling are. A detailed discussion is given of statistical modeling approaches and the feature design taxonomy. The practical relevance of the Poisson distribution assumption for predicting new outlines is explored in [5]

[6] This study highlights the significance of using machine learning techniques such as SGD and SVM in focused crawling. This study emphasizes the importance of keyword optimization for efficient information retrieval during key events. The discussion includes a comparison Among the suggested method with existing approaches in the field of web crawling.[6]

[7] Challenges of data analysis in the e-procurement domain and the need for advanced systems to support decision-making processes. Importance of leveraging past experiences as well as Big Data technologies to improve bid preparation in response to tenders. Potential for extending the system to other application domains and incorporating more recent query expansion techniques.[7]



[8] This study sheds light on the emotional complexities surrounding procrastination and provides insights into demographic variations in sentiment evaluation. The findings highlighted the significance of emotional attitudes in understanding and predicting procrastination behaviors.[8]

[9] This study discusses the significance of context awareness in recommendation systems and highlights The efficiency of deep learning models in capturing complex patterns in user feedback. It also compares the proposed model with existing techniques, emphasizing the advantages of the optimal method using deep recurrent neural networks. [9]

RESULT

[1] The study comes to the conclusion that in order to improve search engine performance and handle the growing complexity of web material, intelligent crawlers must be developed. It emphasizes the importance of continuous research and innovation in crawler technology to meet the evolving demands of web users.[1]

[2]Deep learning models, particularly Bi–LSTM, show promise for POS tagging in Marathi text.

This study highlights the importance deep learning in improving accuracy and efficiency in NLP tasks.

Further research could focus on addressing challenges specific to the Marathi language processing and expanding the dataset for stronger model evaluation.[2]

[3] In conclusion, CAIMANS presents a promising solution for intelligent web artifact crawling and analysis, showcasing advanced technologies and validated usability in real-world scenarios. While demonstrating strengths in scalability and adaptability across diverse domains, the system may face challenges in handling complex web environments and evolving content structures. Despite its potential limitations, CAIMANS offers valuable insights into decision-making processes within organizations, highlighting its potential to enhance strategic planning and information retrieval. Further research and development efforts could focus on addressing weaknesses In order to optimize the system's effectiveness and usability in various industries.[3]

CONCLUSION

[4] Web scraping is a crucial tool in the modern information age, enabling businesses to gather valuable data for analysis and decision-making. As the industry continues to evolve, companies must adapt to stay competitive and meet consumer expectations. The use of advanced technologies such as Python, artificial intelligence, and machine learning can enhance web scraping capabilities. While legal and ethical considerations are important, the benefits of web scraping Regarding data collection and analysis are significant. Embracing web scraping as a strategic tool can provide companies with a competitive edge and valuable insights in today's data-driven market. [4] [5]The study systematically analyzes predictors for new outlines in focused Web crawling and provides valuable insights for improving Web

analytics. The proposed feature design taxonomy and statistical modeling methodology contribute to the understanding of Web change prediction.[5] [7] The intelligent system for focused crawling provided in this document offers a promising solution for organizations to extract valuable insights from Big Data sources, particularly in the e-procurement domain. By leveraging advanced technologies and methodologies, companies can enhance their[7] [8] This study provides valuable insights into emotional attitudes toward procrastination and their predictive role in procrastination predisposition. By uncovering the negative perceptions associated with procrastination and demographic variations in sentiment evaluation, this study offers a nuanced [8] [9] This study successfully introduces a web crawling-based context-aware recommender system using an optimized recurrent neural network in depth. This study demonstrates the efficiency of the suggested model in enhancing recommendation accuracy based on user feedback data. However, Additional investigation is required to address scalability issues and validate the model's performance across diverse recommendation domains. [9]

REFERENCES :

1. "A Survey on Crawlers used in developing Search Engine"
   by Prof. Smita Deshmukh and Kantilal Vishwakarma
2. Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing Dr. Arvind Kiwelekar ,Mrs. Rushali Dhumal (Deshmukh)
3. "An intelligent system for focused crawling from big data sources"
   Ida Bifulco , Stefano Cirillo , Christian Esposito , Roberta Guadagni , Giuseppe Polese
4. "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application" Moaiad Ahmad Khder
5. Look back, look around: A systematic analysis of effective predictors for new outlinks in focused web crawling
   Thi Kim Nhung Danga, Doina Bucura, Berk Atilb, Guillaume Pitelc, d, Frank Ruisa, Hamidreza Kadkhodaeia, Nelly Litvak a, e,∗ a University of Twente.
6. Keyword weight optimization using gradient strategies in event-focused web crawling by Rajiv and Navaneethan
7. [7]Intelligent System for the Focused Crawling of Artifacts from Big Data Sources: A Case Study in the E–Procurement Domain
   Authors: Ida Bifulco, Stefano Cirillo, Christian Esposito, Roberta Guadagni, and Giuseppe Polese
8. Emotional Attitudes toward Procrastination: A Large-Scale Sentiment Analysis .
   Authors: Zhiyi Chen, Rong Zhang, Ting Xu, Yaqi Yang, Junyu Wang, and Tingyong Feng
9. Web–crawling–based Context–Aware Recommender System Using an Optimized Deep Recurrent Neural Network
   Boppana and Sandhya.
10. AI - Based Solution for Web Crawling
    Prashanth Kumar HM1 , Dr. Subramanya Bhat S 2 .
11. AI Classification Cloud-Driven Framework for Web Crawling Using Collective Knowledge by N. Krishnan and Gerard Deepak1.
12. A Survey of Search Engine Optimization for Deep Web Pages ,Jianhui Wang, Junjie Huang, Wei Wu .
13. Automatic Image Captioning for SEO Enhancement: A Deep Learning Approach, Yufeng Zhang, Haoran He, Jianwei Xu, Yifan Jiang, Zhihui He .
14. Towards Understanding the Impact of User Engagement on Web Search Ranking , Yuanzhe He, Yifan Sun, Pengjie Ren, Jiawei Han, Chengzhi Zhang .
15. Deep Learning Based Online Search Personalization with Hierarchical User Profiles , Yifan Sun, Jiawei Han, Chengzhi Zhang, Xiaoting Liu, Jun Xu .
16. Learning to Combine Multiple Features for Effective Web Search Reranking , Yuanzhe He, Jiawei Han, Kang Deng, Xuanhui Huang, Yifan Sun .