A Survey of Focused Crawler Technology: Post-print

Authors: Pan Xiaoying, Chen Liu, Huimin Yu, Zhao Yizhe, Xiao Kangning

Date: 2019-04-01T00:00:00+00:00

Abstract

With the widespread adoption of mobile Internet, network information is growing exponentially, posing significant challenges to effectively extracting and utilizing this information. This paper first introduces the working principles and classification of focused crawlers; then reviews recent research developments on focused crawlers both domestically and internationally, analyzing various topic similarity methods and search strategies, and concludes that compared with general crawler systems, crawler systems based on web content and link analysis achieve substantial improvements in both precision and recall; finally, it analyzes and compares two dynamic search strategies for focused web crawlers and outlines future research directions.

Full Text

Preamble

Vol. 37 No. 5
Application Research of Computers (ChinaXiv Partner Journal)

Survey on Research of Themed Crawling Technique

Pan Xiaoying[1],[2], Chen Liu[1], Yu Huimin[1], Zhao Yizhe[1], Xiao Kangning[1]
([1]School of Computer Science & Technology; [2]Shaanxi Key Laboratory of Network Data Intelligent Processing, Xi' an University of Posts & Telecommunications, Xi' an, Shaanxi 710121, China)

Abstract: With the proliferation of mobile Internet and exponential growth of online information, effectively extracting and utilizing this information presents enormous challenges. This paper first introduces the working principles and classification of topical crawlers. It then reviews recent domestic and international research on topical crawlers, analyzing various topic similarity methods and

search strategies. The analysis reveals that compared to conventional crawler systems, content-based and link-analysis-based crawler systems substantially improve precision and recall rates. Finally, the paper compares two dynamic search strategies for topical web crawlers and discusses future research directions.

Keywords: Web crawler; focused crawler; similarity; Web page content; link analysis
CLC Number: TP393
DOI: 10.19734/j.issn.1001-3695.2018.11.0790

0 Introduction

The Internet represents a massive data repository, with network information resources growing at an exponential rate. Effectively categorizing data as relevant or irrelevant based on user queries and utilizing this information poses a significant challenge for researchers. While commonly used search tools such as Firefox and Google provide general search results, traditional search engines cannot meet modern demands for precise information retrieval. To address the limitations of general-purpose search engines, vertical search engines—directional information retrieval tools—have emerged.

As the core component of vertical search engines, topical crawlers have become an important research direction in the crawler field, attracting widespread attention from researchers worldwide. The key challenge lies in enabling crawlers to capture information more accurately and rapidly. This paper introduces crawler working principles, classification, system architecture, and key technologies, with detailed analysis of both content-based and link-structure-analysis-based topical crawlers. Experimental results demonstrate that compared to general crawler systems, topical crawlers achieve substantial improvements in both precision and recall rates.

1 Working Principle of Web Crawlers

Web crawlers, also known as spiders, automatically browse and retrieve information from the Internet. Search engines rely heavily on web crawlers to traverse the vast sea of online information, capture effective content, and store it for indexing.

The crawling process begins with one or several initial webpages determined by seed URLs. Starting from these seed URLs, the crawler accesses webpages, automatically identifies all embedded URLs, and adds them to a queue of URLs to be crawled. Following a specific search strategy, the crawler visits these queued URLs, downloads the corresponding webpages, stores them in a database, and extracts new URLs from the downloaded content. This process repeats iteratively until the system meets its predefined stopping conditions, at which point

the crawler terminates.

[Figure 1: see original paper] illustrates the implementation principle and process of web crawlers.

2 Classification of Web Crawlers

As shown in [Figure 1: see original paper], the initial URL addresses can be specified manually by users or derived from user-defined domains. Based on implementation technology and system architecture, web crawlers can be categorized into four types: general-purpose Web crawlers, topical Web crawlers, incremental Web crawlers, and deep Web crawlers.

General-purpose crawlers, also known as 全网爬虫 (whole-Web crawlers), target the entire Internet. Beginning with seed URLs, these systems access and collect all hyperlinks from crawled webpages. To prevent duplicate URL acquisition, crawled webpage information is stored in a raw database, parsed, and used to discover new URLs according to the search strategy. This process continues until the collected URLs satisfy the stopping conditions. However, such Web-wide retrieval tools cannot accurately meet specific user needs, prompting the development of topical Web crawlers designed for particular subject requirements. Topical crawlers incorporate additional steps beyond general crawlers: target definition, irrelevant link filtering, and selection of the next URL to crawl.

Topical crawlers perform targeted crawling based on specific themes, focusing on pages relevant to the topic. Initial URLs are obtained through target definition and description. To help crawlers more effectively discover topic-relevant URLs, accurate topic description is essential. The crawler then parses webpage URLs, assesses webpage-topic relevance, predicts link-topic relevance based on search strategies, and determines URL priorities. In focused crawlers, different crawling sequences affect execution efficiency, necessitating search strategies to identify the next URL for crawling and storage. The entire process repeats until the system' s stopping conditions are met.

3 System Architecture of Web Crawlers

Web crawler systems consist of three main modules: webpage acquisition, webpage filtering, and webpage storage. To enable directional extraction of effective information, topical crawlers modify these modules and add a webpage analysis module for calculating webpage similarity, as shown in [Figure 2: see original paper]. The key to topical crawlers lies in defining and describing topics in detail. Before page acquisition, the system evaluates the relevance between webpage text and the topic, enabling the crawler to filter as many topic-relevant pages as possible while excluding irrelevant ones, thereby achieving high accuracy in returned results. Compared to general crawlers, topical crawlers offer several advantages: (a) while general crawlers provide only rough information, topical crawlers have clear themes and can precisely acquire effective information;

(b) topical crawlers must evaluate URL-topic relevance during storage, filtering pages to retain only topic-relevant content.

The main modules of the crawler system are described below:

a) URL Queue. The URL queue stores various hyperlinks, including: the to-be-crawled URL queue for unvisited links; the crawled queue to prevent duplicate page fetching; and the error queue for failed downloads.

b) Webpage Acquisition Module. This module simulates client HTTP requests to Web servers, downloads webpages after receiving server responses, and completes the crawling task. To ensure normal crawler operation and efficiency, a timeout mechanism is implemented—webpages exceeding the time limit are discarded.

c) Webpage Parsing Module. As the central hub connecting other modules, this core component extracts important information links and text from HTML webpages, laying the groundwork for subsequent topic relevance calculations.

d) Webpage Filtering Module. This module filters URLs related to the topic, ensuring the accuracy of the topical crawler system by capturing only relevant pages.

e) Search Scheduling Module. To enable more effective and reasonable URL access, crawlers formulate search rules based on webpage characteristics. Common strategies include depth-first, breadth-first, and best-first. Due to limitations of depth-first, breadth-first and best-first are most commonly used.

f) Webpage Storage Module. This module stores parsed data in files or databases, preparing for search engine retrieval functionality.

g) Preprocessing Module. This module processes webpage content obtained by the parsing module through tokenization, stop-word removal, stemming, etc., converting text into mathematical models recognizable by computers for subsequent similarity calculations.

h) Webpage Analysis Module. As the core of topical crawlers, this module comprises two parts: topic relevance judgment to determine webpage-topic correlation, and topic relevance prediction to forecast URL-topic relevance, enabling prioritized access to topic-relevant URLs through search strategies.

4.1 Webpage Acquisition

The fundamental principle of web crawlers involves simulating browser HTTP requests. The crawler client sends requests to Web servers, downloads webpages after receiving responses, and completes the crawling task.

4.2 Webpage Parsing

Webpage parsing primarily involves webpage denoising—extracting the main content from HTML-structured pages. When extracting webpage content, topical crawlers must analyze HTML structure to retrieve effective information. Common methods include using BeautifulSoup for HTML parsing and employing regular expressions to extract text data.

BeautifulSoup primarily uses XPath and CSS Selector methods to extract required information based on HTML tags, allowing selection via tag, id, class, etc. Chrome and Firefox browsers mark page nodes, enabling direct copying of XPath or CSS Selector paths. Compared to regular expressions, BeautifulSoup is more beginner-friendly. However, for complex page structures, BeautifulSoup is inefficient, requiring fixed page structures with identical tags, ids, and classes for the same fields. Consequently, complex structures necessitate regular expressions, which are complicated but highly efficient for extracting string-structured information.

4.3 Data Storage

Crawled data is typically stored either locally in CSV/Excel formats or directly in databases. Small datasets can be saved locally, while large-scale crawlers generally use databases for convenient storage and subsequent analysis. Python' s built-in CSV package facilitates writing to CSV or Excel tables during crawling. Database storage includes relational databases (MySQL, SQL Server) and non-relational databases (MongoDB, SSDB, HBase).

There are two approaches to database writing: (1) centralized vectorized cleaning and bulk import after all data is crawled; (2) incremental cleaning and import during crawling. For large-scale crawlers, stability is crucial. Network errors inevitably occur during lengthy crawling processes, rendering the first approach' s data useless upon failure. The second approach avoids such issues, with faster single-pass cleaning and import that doesn't impact overall efficiency, making it the preferred method.

4.4 Topic Discrimination

Topic discrimination primarily assesses crawled webpage-topic relevance, beginning with topic definition. This problem is often explored as text classification. Researchers currently combine anchor text, webpage tags, etc., to calculate URL-topic relevance, making topic relevance calculation a key differentiator among topical crawlers. Common similarity algorithms include the Vector Space Model (VSM) and semantic similarity.

1) Vector Space Model. This conceptually simple model converts text processing into vector operations in vector space, representing each document as a dimension and measuring similarity through spatial vector calculations.

2) Semantic Similarity. Unlike English, Chinese descriptions of objects can vary significantly. Semantic understanding in natural language processing has long perplexed researchers. Traditional tokenization and term frequency statistics cannot accurately comprehend textual meaning, reducing recognition accuracy. Observable quantities are limited to term frequency and document frequency, which form the basis for semantic analysis methods enabling computers to "understand" human language.

4.5 Webpage Search Strategies

As directional crawlers with specific themes, topical crawlers aim to quickly and accurately search for topic-relevant pages. Search strategies ensure orderly, purposeful crawling, enabling efficient task completion through reasonable path selection.

Search strategies are categorized as static or dynamic based on whether rules are predetermined. Static strategies follow fixed rules unaffected by webpage structure or content changes. Dynamic strategies prioritize efficiency and speed, adjusting crawling routes in real-time. The Internet comprises webpages and hyperlinks; dynamic strategies can be content-based or link-analysis-based.

Different webpage content reflects different meanings—titles, keywords, and text are most representative. Dynamic strategies require rapid link relevance calculation, making local-text-based strategies common due to their low computational cost. Global-text-based strategies using all webpage text are time-consuming. Classic content-based strategies include Fish-Search and Shark-Search. Link-analysis-based strategies rest on three principles: (a) webpage value is proportional to citations; (b) cited webpages exhibit greater structural and content similarity; (c) well-structured pages are easily cited. These strategies analyze and predict webpage topics using links to evaluate URL priorities. Classic approaches include PageRank, HITS, and HillTop.

5 Research Directions for Topical Crawlers

In recent years, researchers have developed crawling strategies and algorithms to improve topical crawler accuracy and efficiency, primarily focusing on topic similarity and search strategies. Current research is divided into several areas:

5.1 Content-Based Topical Crawlers

Different webpage content reflects different meanings, with titles, keywords, and text being most representative. Wang Jinyang [**?**] proposed using titles to construct concise content subtrees for topic judgment, employing semantic similarity to modify VSM for relevance determination. This approach addresses traditional VSM' s lack of semantic judgment, improving topic recognition accuracy and crawler precision.

Zhou Mixue [?] designed a medical vertical search engine using topical crawlers, evaluating relevance from hyperlinks, meta-information, and lexicons to effectively filter topic-relevant pages. To address traditional PageRank limitations, the study introduced temporal feedback, authority, and topic relevance factors, significantly improving medical vertical search precision.

Li Hongzhi et al. [?] constructed a KNN classifier for webpage-topic relevance using IK Analyzer for Chinese word segmentation and TF-IDF for feature extraction. Results showed that KNN-based crawlers achieved higher accuracy with more documents, outperforming traditional PageRank and Bayes algorithms in classification effectiveness and stability.

Zhang Lijing et al. [?] applied topical crawlers to book themes, designing the ODP2EVSM algorithm. This approach comprises two parts: dynamic keyword expansion using the Open Directory Project (ODP) for accurate topic description, and VSM-based term semantic extension for relevance judgment.

Li Hui et al. [?] used VSM to calculate content similarity, enabling effective filtering of highly topic-relevant pages while improving crawling efficiency and accuracy. Applied to aquaculture input quality supervision, tests demonstrated stable operation and high information acquisition accuracy.

Ji Xiang [?] obtained an SVM classifier from agricultural product price samples, then constructed a KNN classifier using support vectors for effective page classification. To accurately and efficiently collect all agricultural product price information, the system employed both SVM and support-vector KNN classifiers under different conditions.

To address polysemy in webpage text, Meng Zhu [?] proposed using semantic models combined with pointwise mutual information to determine word meanings based on context, jointly evaluating webpage links for topic relevance. Wang [?] expanded professional vocabulary (e.g., electronics brands) into custom dictionaries, significantly improving query accuracy by modifying the Heritrix framework for an electronics search engine.

Song et al. [?] introduced a dynamic topical crawler system based on keywords and SVM models, effectively acquiring target information for applications in information security and corporate crisis management. Dahiwale et al. [?] proposed a semantic-focused Web crawler using Meta tags as the primary information source for relevance calculation before page download, improving document quality by filtering irrelevant links.

Topical crawler research has developed various algorithms for similarity judgment, involving text similarity assessment. Current methods fall into two categories: statistical models like VSM, and semantic understanding models. Researchers aim to leverage semantic relevance for more precise results. The process involves defining crawler topics, calculating webpage and URL relevance based on content and structure, and determining crawling priorities. Such crawlers typically achieve high accuracy.

presents experimental data from content-based topical crawler algorithms. Performance metrics include precision, recall, and F-value. Results demonstrate that detailed webpage content significantly improves crawler precision and recall.

5.2 Link-Analysis-Based Topical Crawlers

Traditional content-based evaluation strategies often overlook link correlations, while link-analysis-based strategies neglect webpage content, causing "topic drift." Cai Guangbo [**?**] combined the content-based Fish-Search algorithm with the link-analysis-based PageRank algorithm, integrating webpage text and links to calculate page-topic relevance, significantly improving precision.

Hu Pingrui et al. [**?**] proposed a URL-pattern-set-based topical crawler leveraging URL structural and semantic similarity features, which vary significantly across modules. By distinguishing URL feature differences, the method predicts URL priorities based on pattern importance, ensuring high precision and recall.

Zhang Jin et al. [**?**] introduced a link-ranking algorithm based on sub-link analysis to ensure high topic relevance. By weighting current link relevance with sub-link relevance and scoring links accordingly, the algorithm improves crawling accuracy.

Shi Baoming et al. [**?**] proposed a link-model-based relevance discrimination algorithm calculating URL-topic relevance, demonstrating higher efficiency than traditional methods.

Liu et al. [**?**] employed the VIPS algorithm to analyze webpage depth, using multi-granularity Shark-Search combined with query-based hit algorithms to improve crawling strategy. This new algorithm remedied shortcomings of both Shark and HITS, reducing noise and eliminating topic drift.

Kumar et al. [**?**] utilized anchor text richness, building a page analyzer to understand content and anchor context for crawling decisions, guiding crawlers in specific domains.

Liu Shaotao et al. [**?**] combined content-based Best-First with HITS algorithms, designing a new link selection strategy that integrates content and link structure to improve topic relevance and authority during downloading.

Pant et al. [**?**] studied crawler navigation using link contexts to predict hyperlink advantages relative to starting topics, examining various link context definitions' impact on performance using SVM-guided topical crawlers.

Shen et al. [**?**] proposed a complex-network community-based crawling method, dividing the process into community detection for link structure analysis and topic-relevant analysis within identified communities.

Gupta et al. [**?**] extracted link context using tag tree and LALR parsing methods, with the tag tree identifying anchor text concepts for context extraction.

Peng et al. [?] argued that anchor text may inadequately express webpage meaning, potentially misleading crawlers. They proposed dividing pages into smaller regions to avoid obscuring highly relevant areas, selecting link context based on regional relevance.

Geng et al. [?] improved crawler efficiency and accuracy by combining HTML analysis with text density for extraction and incorporating multi-factor similarity calculation (treating news text differently), significantly improving text extraction accuracy.

Shark-Search performs well near relevant pages but lacks a "global view," while PageRank's iterative nature increases weights in tightly connected regions, causing topic drift. Qiu et al. [?] merged Shark-Search with PageRank, using the former for webpage scoring and the latter for URL link weighting to define page importance, remedying both algorithms' defects. Results show suitability for large-scale page collection.

With billions of webpages linked via hyperlinks, researchers aim to effectively extract link context meaning through parsing and extraction or by improving traditional link selection algorithms based on webpage content. These approaches analyze links to determine importance, emphasize link authority' s significance to user needs, and combine content and link analysis to solve topic drift and improve accuracy.

presents experimental data from link-analysis-based topical crawler algorithms. Results demonstrate that link-analysis-based crawlers compensate for content-based crawlers' neglect of sub-link impacts, with combined approaches yielding more precise collection results.

6 Applications of Crawler Systems in Various Fields

With exponential information growth, vertical search engines for specific domains have become research hotspots, spawning domain-specific topical crawlers.

In smart agriculture and forestry, Zhang Lulu [?] designed a pest-themed search engine using domain lexicons for detailed topic description, integrating website links and content. Li Hui et al. [?] employed topical crawlers as a key step in aquaculture input quality supervision systems, avoiding irrelevant page downloads and improving precision and recall. Meng Fanjiang et al. [?] built an agricultural product price search engine that plays a crucial role in collecting price data and identifying change factors.

In healthcare, Yin Man [?] constructed a medical equipment vertical search engine analyzing product characteristics and stakeholder needs. Zhou Mixue [?] improved medical vertical search precision through topic similarity and PageRank enhancements. Li Xuebo [?] designed a traditional Chinese medicine topical crawler for reliable, precise health information services.

In education, Liu Can et al. [?] applied topical crawler technology for personalized educational news recommendation. Hu R et al. [?] developed an efficient focused crawler using the MEAN stack (MongoDB + Express + AngularJS + Node.js) with Cheerio, providing substantial effective data. Li Xiangyu [?] developed a biosafety domain crawler for precise knowledge acquisition, while Guan Weiguo [?] collected food contact material safety information for network public opinion monitoring.

7 Development Trends of Topical Crawlers

Despite extensive research, topical crawler performance offers substantial room for improvement:

a) Adaptive Search Strategies. Current crawlers use fixed strategies, but varying website structures require integrated crawling rules for enhanced performance.

b) Semantic Understanding for Fine-Grained Topics. While broad topics benefit from content and link context, fine-grained topics face limitations like inaccurate keyword description, reducing precision and recall. Semantic improvements in feature selection represent a future research hotspot.

c) Anti-Crawling Measures. Websites implement anti-crawling strategies for information protection. While distributed crawlers address this, their high development cost raises questions about designing cost-effective advanced crawlers.

d) Temporal Dynamics in Hot Topics. Traditional methods struggle to describe hot topics accurately. Leveraging temporal characteristics—generation, development, and dissipation times—could enable real-time tracking, such as in food safety emergency topic detection.

References

[1] Wang Jinyang. Parallelization of thematic Web crawlers [D]. Chengdu: Southwest Petroleum University, 2017.

[2] Peng Xiaoming. Design and implementation of the theme crawler [D]. Beijing: Beijing University of Posts and Telecommunications, 2013.

[3] Wang Congrui. Research on key technologies of subject reptiles [D]. Shijiazhuang: Shijiazhuang Railway University, 2015.

[4] Zhou Mixue, Research and implementation of medical vertical search engine based on improved PageRank algorithm [D]. Xi' an: Chang' an University. 2017.

[5] Li Hongzhi, Song Jie. Thematic Web crawler based on knn classification algorithm [J]. Journal of Yibin University, 2017, 17 (12): 61-65.

[6] Zhang Lizhen, Zeng Qingtao, Li Yeli, et al. Research on crawling algorithm for book theme [J]. Journal of Computer Science, 2017, 44 (b11): 460-463.

[7] Li Hui, Zhang Biao, Wu Wenliang. Quality information supervision system for aquaculture inputs based on subject reptile algorithm [J]. Jiangsu Agricultural Sciences, 2017, 45 (8): 210-214.

[8] Ji Xiang. Research and implementation of agricultural product price theme search engine [D]. Harbin: Northeast Agricultural University, 2017.

[9] Meng Zhu. Research on semantic model of word vector and its application in subject reptile system [D]. Beijing: China University of Geosciences, 2017.

[10] Wang Aihua. Design and implementation of vertical search platform for electronic product information [C]// Proc of International Conference on Robots & Intelligent System. Washington DC: IEEE Computer Society, 2017: 101-104.

[11] Song Biao, Zhu Jianming, Zhang Jianguang. A research of dynamic theme crawler based on keywords and support vector machine [C]// Proc of International Conference on Management Science & Engineering. San Francisco: IEEE Press, 2014: 21-26.

[12] Dahiwale P, Raghuwanshi M M, Malik L. Design of improved focused Web crawler by analyzing semantic nature of URL and anchor text [C]// Proc of International Conference on Industrial and Information Systems. San Francisco: IEEE Press, 2015: 1-6.

[13] Cai Guangbo. Design and implementation of topic-oriented multi-threaded web crawler [D]. Lanzhou: Northwest University for Nationalities, 2017.

[14] Hu Pingrui, Li Shijun. Theme crawler based on URL pattern set [J]. Journal of Computer Applications, 2018, 35 (3): 694-726.

[15] Zhang Jin, Ni Xiaojun. Research on topic crawling strategy based on semantic tree and VSM [J]. Computer Technology and Development, 2017, 27 (11): 66-70.

[16] Shi Baoming, He Yuanxiang, Wu Chongzheng. Research on crawler search strategy in topic search engine [J]. Computer Engineering and Applications, 2014, 50 (2): 116-119.

[17] Liu Naiwen, Yao Rongbao. The crawling strategy of shark-search algorithm based on multi granularity [C]// Proc of International Symposium on Computational Intelligence and Design. San Francisco: IEEE Press, 2016: 41-44.

[18] Kumar N, Singh M. Framework for distributed semantic Web crawler [C]// Proc of International Conference on Computational Intelligence and Communication Networks. San Francisco: IEEE Press, 2016: 67-71.

[19] Liu Yutao, Li Hongsheng. The theme reptile algorithm based on fusion link structure [J]. Journal of Huaqiao University :Natural Science , 2017, 38 (2): 195-200.

[20] Pant G, Srinivasan P. Link contexts in classifier-guided topical crawlers [J]. IEEE Trans on Knowledge & Data Engineering, 2005, 18 (1): 127-136.

[21] Shen Guilan, Sun Jie, Yang Xiaoping. A focused crawling method based on detecting communities in complex networks [J]. Journal of Henan Normal University, 2014, 9 (8): 187-196.

[22] Gupta S, Yadav S. Extraction of link context using tag tree and LALR parsing [C]// Proc of Information & Communication Technologies. San Francisco: IEEE Press, 2013: 253-257.

[23] Peng Tao, Liu Lu. Focused crawling enhanced by CBP-SLC [J]. Knowledge-Based Systems, 2013, 51 (1): 15-26.

[24] Geng Zhingqiang, Shang Dirui, Zhu Qunxiong, et al. Research on improved focused crawler and its application in food safety public opinion analysis [C]// Proc of Chinese Automation Congress. Beijing, 2017: 2847-2852.

[25] Qiu Lei, Lou Yuansheng, Chang Ming. Research on theme crawler based on shark-search and PageRank algorithm [C]// Proc of International Conference on Cloud Computing and Intelligence Systems. San Francisco: IEEE Press, 2016: 268-271.

[26] Xiao Jiang, Ji Jie. The application of focused crawler based on Heritrix in internet public opinion system [J]. Electronic Design Engineering, 2015, 23 (6): 29-31.

[27] Zhang Lulu. Research on pest and disease subject search engine based on distributed acquisition strategy [D]. Harbin: Northeast Forestry University, 2017.

[28] Meng Fanjiang, Ji Xiang, Yuan Qi, et al. Research and implementation of agricultural product price subject search engine [J]. Journal of Northeast Agricultural University, 2016, 47 (9): 64-71.

[29] Yin Man. Design and implementation of medical equipment vertical search engine [D]. Chongqing: Chongqing University. 2017.

[30] Li Xuebo. Research on Chinese medicine web information resource evaluation system based on hadoop [D]. Jinan: Shandong Medical University, 2016.

[31] Liu Can, Ren Jianyu, Li Wei, et al. Education news crawling and display system for personalized recommendations [J]. Software Engineering, 2018, 21 (2): 34-40.

[32] Hu Rong, Feng Zhongke, Jiang Junzhiwei. Web crawler of atmosphere and weather data based on MEAN stack with CheerIO [J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47 (6): 323-328.

[33] Li Xiangyu. Ontology modeling and knowledge platform development in biosafety field [D]. Tianjin: Tianjin University, 2016.

[34] Guan Weiguo. Design and implementation of reptile system for food contact material safety [D]. Shanghai: Donghua University, 2017.

[35] Ding Shenchun, Gong Silan, Zhou Wenjie, et al. Research on network public opinion real-time monitoring of the south china sea issue based on knowledge base and focused crawler [J]. Journal of Intelligence, 2016, 35 (5): 34-37.

Note: Figure translations are in progress. See original paper for figures.