Visualization Analysis of Web Crawler Evolution Retrieval Research Based on KG

Zhenjiang Chen[1,2],Jiamian Wang[3], Yizhun Peng[1,2*],Di Li[1,2],Lianchen Zhao[1,2]
[1]*College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, 300222, China*
[2]*Advanced Structural Integrity International Joint Research Centre, Tianjin University of Science and Technology, Tianjin, 300222, China*
[3]*College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, 300222, China)*
** pengyizhun@tust.edu.cn*

Abstract

In order to understand the basic situation and future development trend of domestic research on web crawler technology. By using Citespace information visualization analysis software, 2892 web crawler technical literatures in CNKI information technology database from 2000 to 2018 were data mining. From the aspects of literature time distribution, inter-agency cooperation network analysis, co-citation of authors, co-occurrence of keywords and analysis of research frontiers, this paper draws a map of scientific knowledge and sorts out the research background. This paper intuitively reveals the research status, development path, core research groups and research fields of web crawler technology.
Keywords: web crawler; knowledge graph; CiteSpace; co-occurrence analysis; visualization

1. Introduction

Maybe for many people, "Web Crawler" sounds like a new concept like "big data" or "machine learning". But in fact, web crawlers have a long history, which can be traced back to the birth of the world wide web. At that time, the Internet had no search function. The Internet is just a collection of file transfer protocol (FTP) sites. Users can only browse certain websites to find specific shared files. So developers created an automated program called a web crawler or search engine robot. It can grab all the web pages on the Internet, and then copy the contents of all the pages to the database for indexing.

Web crawler technology is the first step to obtain massive network text resources, is an important acquisition technology tool, and is also an important topic in the field of data information. With the rapid development of the information age, network crawler technology has been continuously developed and improved, which has attracted widespread attention from emerging enterprises and national information units. In order to better show and explain the overall framework of research on web crawler technology, so that domestic scholars can quickly understand its research status and trends. In this paper, the research of web crawler technology is deeply analyzed by using the CNKI information technology database scientific literature and the visualization software CiteSpace of scientific knowledge graph.[1]

2. Data Sources and Analysis Tools

2.1. *Data sources*

CNKI is an international leading network publishing platform integrating journals, doctoral dissertations, master's dissertations, conference papers, newspapers, reference books, yearbooks, patents, standards, traditional Chinese studies and overseas literature resources. The data source of this paper is 2892 articles about describing web crawler technology from 2000 to 2018 in the information technology database of CNKI. On this basis, bibliometric statistics and information mining are carried out. This ensures the relative scientificity of the research results. As shown in Table 1.

Table 1. Data acquisition content

Category	Description
Data Source	CNKI China HowNet Database
Retrieval condition	topic='Web crawler'or'research'
Time limit	2000-2018
Document Category	Journal Papers & Doctoral Dissertations
Number of valid papers	2,892

2.2. *Analysis Tool*

CiteSpace is a visual analysis tool based on Java platform developed by Professor Chen Chaomei of Drexel University. Since it was put forward, it has been widely used in various fields because of its scientificity and universality.[2] In this paper, the author's co-occurrence analysis, mechanism co-occurrence analysis and keyword co-occurrence analysis of 2,892 papers are carried out by CiteSpace. Through clustering view and time zone view, we can draw a scientific knowledge graph of the current situation, hot spots and frontiers of network crawler research field.[3]

3. Results and Analysis

3.1. *Time Distribution of Selected Literatures*

In CNKI, 2892 articles were retrieved by topic retrieval, and the annual trend distribution map of research literature on web crawler technology was obtained by preliminary sorting out.

As shown in Figure 1, the number of CNKI papers published continues to rise, from zero in 2000 and 2001 to two in 2003, achieving zero breakthroughs in China. According to the analysis, in 2000, Salesforce and eBay launched their own API, which can be used by programmers to access and download some public data. Since then, many websites have provided Web APIs to enable people to access their public databases. Therefore, the technology of web crawler has gradually developed. In 2004, Python Beautiful Soup could grab some data without the API provided by the website. Since then, the number of CNKI papers has risen steadily, reaching its peak in 2017, and the number of CNKI papers has reached 356. With the continuous development of network technology, programmers are still developing a way to improve the network crawler to adapt to the expanding network data.

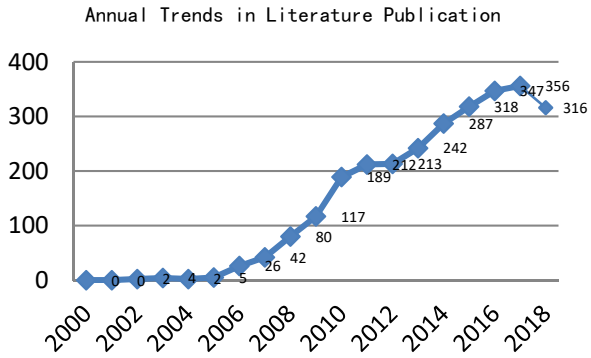

Fig. 1. Annual Trend Distribution of Web Crawler Technology Literature Publication

3.2. Analysis of Cooperation Network of Domestic Research Institutions

Introduce the data into CiteSpace software, select "Institution" as the node, select Top N = 50 as the threshold, set the value of "Years Per Slice" to 3, and select 20% of the data before 3 years for clustering. Run CiteSpace to get the knowledge map of the subordinate organizations of journal papers of web crawlers, and then get Figure 2 and Table 2 after sorting out. A node in the atlas represents an organization unit, and the size of the node represents the number of documents sent by the organization unit. Mediation centrality is an index to measure the importance of nodes in the whole network atlas, which can indicate the importance of documents in the database in CiteSpace visualization software. The greater the centrality of intermediary, the greater its influence in this field.[4]

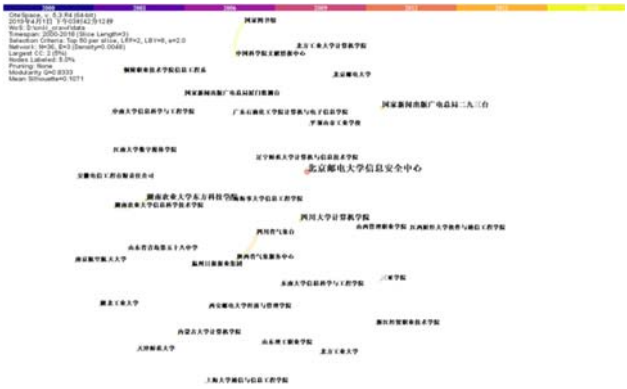

Fig. 2. Knowledge Graph of Publishing Agencies

By analyzing the statistics of the publishing agencies, we can get the universities and institutes

which have more research on Web crawlers. As shown in the figure above, the top four in the research of web crawlers in China are the Information Security Center of Beijing University of Posts and Telecommunications, the Computer College of Sichuan University, the Oriental Institute of Science and Technology of Hunan Agricultural University and the 293 National Bureau of Press, Publication, Radio and Television. In terms of the number of articles published, although there are many institutions in our country, they are not prominent, and the overall academic strength is relatively weak. This also shows that more scholars are still needed to improve our country's development in this field.

Table 2. Top 6 in the Number of Institutional Publications

Institutional	Year	Quantity
Information Security Center of Beijing University of Posts and Telecommunications	2010	5
School of Oriental Science and Technology, Hunan Agricultural University	2016	3
School of Computer Science, Sichuan University	2016	3
State Administration of Press, Publication, Radio and Television 293	2016	3
Wenzhou Daily Newspaper Group	2016	2
Nanjing University of Aeronautics and Astronautics	2016	2

3.3. Co-citation Analysis of Authors

To a certain extent, the number of papers published by the author can roughly reflect the scientific research ability and level of the scholar in the relevant fields, and indirectly reflect the research maturity in the corresponding fields.

Based on the data in CNKI's information technology database, the authors of network crawler technology research are analyzed in time domain. Set the value of "Years Per Slice" to 1, and the selection criterion of node type to "Top N=50". Run CiteSpace to get the distribution time zone map of the important authors in the domestic network crawler research. As shown in Figure 3. Table 3 lists the core researchers of 15 authors who have published more articles, such as Liu Qiang, Zhou Ping and Li Baoguo.

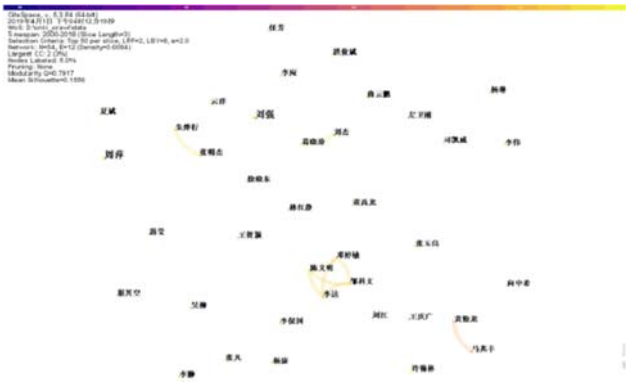

Fig. 3. Knowledge Graph of Publishers

Table 3. Top 10 Authors of Internet Crawler Research in China

Author's Name	Sending the papers amount	Author's Name	Sending the papers amount
Liu Qiang	3	Ren Fang	2
Zhou Ping	3	Chen Yiming	2
Li Baoguo	2	Yunyang	2
Li Ying	2	Xiang Zhongxi	2
Hong Junbin	2	Zhang Yugao	2

3.4. High Frequency Keyword Statistics and Co-occurrence Clustering Analysis

High frequency keywords can be used to analyze the research hotspots of scientific inquiry. Using CiteSpace software, the high-frequency keywords of 2,892 research papers from 2000 to 2018 are counted and visualized knowledge maps are drawn. Table 4 lists the top 10 and the middle 10 (parts) and the last 10 high-frequency keywords, years, frequencies and intermediary centrality (abbreviated as "neutrality").

Table 4. Keyword Distribution of Web Crawler Related Literature (Part)

Frequency	Year	Intermediateness	Key word
713	2004	0.21	Internet worm
238	2006	0.12	Search Engines
135	2007	0.17	Topical crawler
122	2009	0.13	Reptile
101	2008	0.09	Vertical Search Engine
96	2008	0.13	Internet public opinion
85	2008	0.19	Text classification
79	2008	0.11	Chinese word segmentation
78	2010	0.11	Data mining

73	2008	0.07	Information extraction
37	2006	0.08	Information Retrieval
35	2012	0.05	Text clustering
35	2010	0.07	Distributed Reptilesl
33	2015	0.02	Big data
32	2007	0.07	Vector Space Mode
30	2010	0.01	Text Mining
29	2015	0.01	Scrapy
29	2011	0.05	Machine learning
27	2009	0.04	Ajax
26	2011	0.03	Topical crawler
6	2013	0	Web Services
5	2016	0	User Behavior
5	2017	0	Word vector
4	2013	0	Visualization
2	2013	0	Web Page Clustering
2	2015	0	Management information systems
2	2015	0	Task scheduling
2	2014	0	Tendency analysis
2	2015	0	Doctor's recommendation
2	2010	0	Web Page Purification

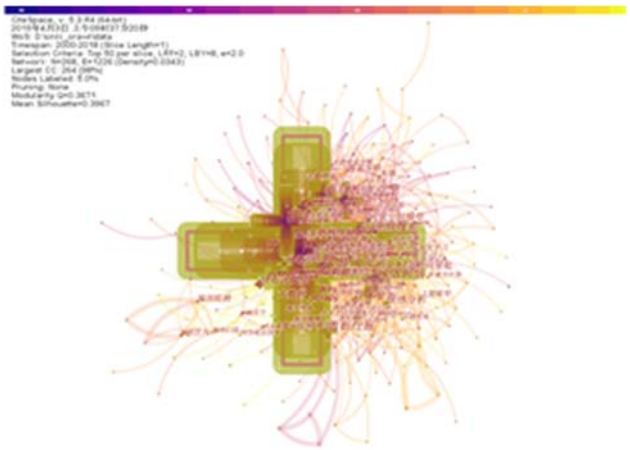

Fig. 4. Keyword Co-occurrence Map in the Field of Web Crawler Technology

The keywords with high frequency and high intermediary centrality represent the new research frontiers in the corresponding time interval in the field of network crawler research to a certain extent, and their diachronic evolution reflects the dynamic changes of hot spots in a discipline or research field.

Combining with the emergence of Fig. 4, Fig. 5 and Table 4, we can see that the research hotspots of web crawlers maybe: web crawler--vertical search engine--topic crawler--text categorization--distributed crawler--vector space model--machine earning--deep learning--knowledge graph (knowledge base). Of course, some parts are still in-depth research, such as topic crawler research will focus on how to improve the accuracy of web page correlation calculation, how to reduce the space-time complexity of calculation, and how to enhance the adaptability of crawler.[5] At the same time, it can be found that network data mining based on large data crawler technology provides basic data support for machine learning, deep learning and knowledge atlas.[6]

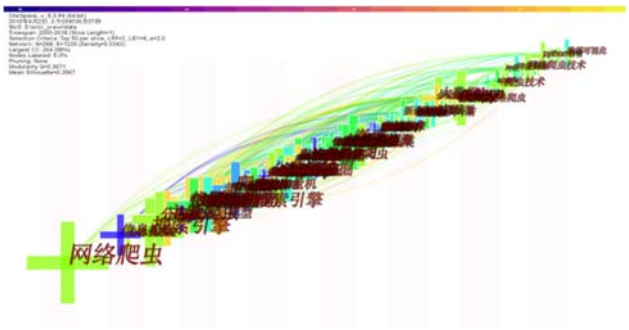

Fig. 5. Keyword Sequence Map Atlas

As shown in the figure above, the cold and warm colors of the graph indicate the relationship between time and time, that is, the colder the color is, the farther the research content is from now, and the warmer the color is, the trend of current research is indicated.

3.5. Research Frontier Analysis

In the era of big data, data sources are the first thing to do data analysis, and the network resources are the most abundant. So we often need to do network crawlers, which can let us get more data sources, and these data sources can be collected according to our purposes, removing a lot of irrelevant data.

Web crawler technology has now formed a relatively complete technical system, different types of Web sites for different forms of data capture. Then further processing of the captured data, such as data cleaning, is the reprocessing of the captured results, which can effectively improve the quality of the captured data, especially in the heterogeneous data source environment of the World Wide Web.[7] Secondly, different types of databases are used to store different types of data, such as relational databases and non-relational databases. Then the valid data is processed. Text mining, for example, is an

interdisciplinary subject, involving data mining, machine learning, pattern recognition, artificial intelligence, statistics, computer linguistics, computer network technology, informatics and other fields.[8] It is a method and tool to discover implicit knowledge and patterns from many documents. It is developed from data mining, but it is different from traditional data mining.[9] Machine Learning, for example, refers to the process of using some algorithms to guide computers to use known data to obtain appropriate models, and to use this model to give judgments on new situations.[10] The idea of machine learning is not complicated, it is only a simulation of the learning process in human life, and in the whole process, the most critical is data.

4. Conclusions

This paper analyses the amount of literature, source journals, authors and organizations, keyword analysis and hot spot analysis of web crawlers, and draws the following conclusions:

4.1. Further research is needed

From the point of view of research basis, with the continuous development of the era of big data, there are 2892 core journals on Web crawler, and the research in the field of web crawler technology has begun to be saturated. However, the evolution of machine learning, in-depth learning and knowledge mapping needs further research and more attention.

4.2. Institutions and authors

Although there are many institutions in our country to study the technology of web crawler, they are less cited, the overall academic strength is relatively weak, and the communication between authors is not very close, and the total cited number of authors is also very small, indicating that more scholars are still needed to step into and improve the field.

4.3. High frequency keyword statistics and co-occurrence cluster analysis

Through keyword co-occurrence analysis, we can see that the research on the evolution of web crawler can be divided into three levels: basic crawler technology level, search crawler level and application level of crawler. Basic crawler technology includes subject crawler, information search, text categorization and distributed crawler. Search crawler level includes vertical search engine, information retrieval and subject retrieval. The application level of crawler includes text mining, machine learning, in-depth learning, personalized recommendation, natural language processing and knowledge graph.

4.4. Research hotspots and frontiers

Through the analysis of the research frontier, it is found that the research frontier of web crawler technology is mainly in the fields of data visualization, machine learning, in-depth learning and knowledge atlas.

References

1. Li Jie, Chen Chaomei. Cite Space: Technological Text Mining and Visualization [M]. Beijing: Capital University of Economics and Trade Press, 2016:32.
2. Li Jie. CiteSpace Chinese Guide [M/OL] http://blog.sciencenet.cn/blog-554179-1066981.html.
3. Chen Chaomei. CiteSpace Ⅱ. Detecting and Visualizing Emerging Trends and Transient Patterns in ScientificLiteratur[J]. Journalof the American Society for Information Science and Technology,2006(3).10.
4. Lin Deming, Chen Chaomei and Liu Zeyuan. A study of Zipf-Pareto distribution of network intermediary centrality [J]. Informatics, 2011, 30 (1): 76-82.
5. Zhong Weijin, Li Jia. Co-word analysis: the process and method of Co-word Analysis [J]. Information Journal, 2008, 27 (5): 70-72.
6. Pan Xiaoying, Chen Liu, Yu Huimin, Zhao Yizhao, Xiao Kangzhen. Summary of research on topic crawler technology [J/OL]. Computer application research: 1-6 [2019-04-02].https://doi.org/10.19734/j.issn.1001-3695.2018.11.0790.
7. Zhou Lizhu, Lin Ling. Review of focused crawler technology [J].Computer Applications, 2005 (09): 1965-1969.
8. Wang Minxiang. Research and implementation of topic-oriented crawling search strategy [D]. Shaanxi Normal University, 2012.
9. Feng Li.Overview of Reptilian Technology[J].Computer Knowledge and Technology, 2017,13(27): 213-214.
10. Shi Hongyi. Overview of Machine Learning [J]. Communication World, 2018 (10): 253-254.