Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1932–1939, DOI 10.12694/scpe.v25i3.2794

HIGH-PERFORMANCE COMPUTING WEB SEARCH SYSTEM BASED ON COMPUTER BIG DATA

YINGXI KANG[*] BEIPING TANG, AND XIAODONG HU

Abstract. File sharing, streaming media, collaborative computing, and other P2P systems are all unicast to establish the corresponding overlapping network. The superimposed network is generally carried out based on the existing primary network. In this way, the access of each node is random. At the same time, this will cause the topological structure of the upper and lower layers to be inconsistent. This will increase the communication delay between nodes and cause an excellent bandwidth burden to the underlying network. The existing topology matching methods still face problems, such as poor scalability and long node aggregation time. This paper aims to design a topological distributed node aggregation method based on network coordination and distributed hash table (DHT) algorithm. This paper established a two-dimensional mesh model of nodes based on equal-distance concentric circles and divided into two equal areas. The parts of multiple namespaces correspond one by one according to their location. Because nodes are kept close, neighbours can be aggregated through DHT's primary "publish" and "search" primitives. Experimental results show that the TANRA method can match the network's topology under a slight delay and a large number of nodes. The TANRA method can effectively reduce the path delay in structured networks.

Key words: Topological induction; Node proximity; Topology aware node aggregation algorithm; Node cluster; Distributed hash table; Overlay Network

1. Introduction. Peer-to-peer (P2P) technology has been widely concerned with its excellent scalability and fault tolerance. P2P networks can make full use of the node resources of the network edge system. It is getting more and more attention in the new wave of applications on the Internet. Especially in mass content publishing and streaming media, the resource sharing of intermediate nodes can effectively reduce the consumption of network resources where the data source resides [1]. Therefore, highly concurrent content distribution and media transmission can be achieved. In P2P mode, there are a series of streaming media technologies such as PROMISE and Cool streaming. Although the above systems have good scalability and support ability for high concurrency services, the lack of corresponding topological identification methods causes the topological structure of the upper-layer overlay network and the lower-layer communication network to be inconsistent. In this case, the nodes in the network cannot exchange information, which makes the data transmission efficiency in the high-level network low and affects the overall performance. At the same time, it also brings a lot of bandwidth burden to the underlying physical network. Studies show that P2P has accounted for 60% of the business on the Internet in recent years. Therefore, reducing the occupation of P2P networks under the premise of ensuring the quality of network service is an urgent problem that needs to be studied [2]. Therefore, a highly scalable node aggregation algorithm is designed in this paper. Then, a topologically aware distributed node aggregation algorithm (TANRA) based on the network coordinate algorithm and DHT algorithm is proposed. Unlike other methods, the proposed method can achieve topology consistency between the global and underlying networks. At the same time, it can converge quickly and increase the overhead of nodes. This scheme is suitable for many node configurations in P2P networks.

2. HPC web search system framework. The architecture of the proposed in-site search system is shown in Figure 2.1 (The picture is quoted from Example of system architecture picture in TikZ). The system consists of two parts: index sub and search sub. It mainly completes the index creation, increment, update, delete, and other functions in the Xapian database and queries it according to users' requirements. Interact with the Xapian database through a web interface. Xapian has a separate index database. It is independent

[*] Hunan Institute of Engineering, Xiangtan, 411104 Hunan, China (Corresponding author, 06158@hnie.edu.cn)

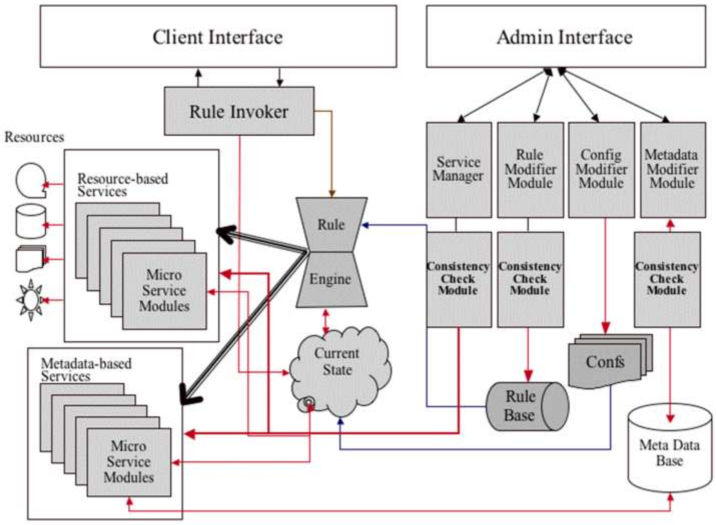

Fig. 2.1: Architecture diagram of the site finding system.

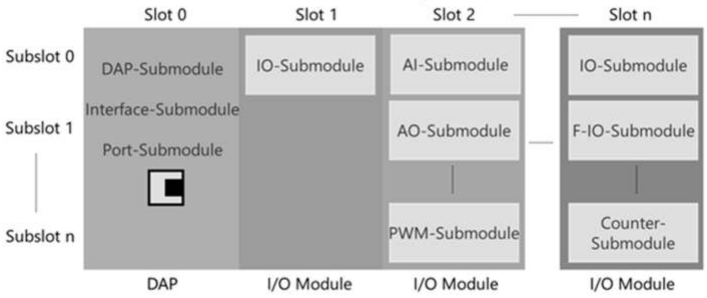

Fig. 2.2: Schematic diagram of index submodule.

of the site's database [3]. This is beneficial for strengthening the independence and flexibility of the Xapian website. An effective large-scale data query system is established based on adequate information query.

2.1. Index submodule. A lot of new data is generated on the site every day. The metrics are not the same. These metric changes when the site is upgraded. To improve the running speed of the system and improve the user experience of the system, the author also designed a site-independent index submodule. The component's architecture is shown in Figure 2.2 (the image is referenced).

The index is created to read the Web page data from the site's database. For the title of the page and the content, call the segmentation module, and then get the segmentation words and create a Xapian file for each page. Each Xapian file has a unique file ID number [4]. Store the split word as a Term in this file. In addition, the Values structure of the file also retains related data such as city ID number and user name so that the file can be easily and quickly filtered, sorted, and deleted duplicate items. This is the relay data attached to the file. Its purpose is fast access to matches, matching, and filtering. Some data in the file can be used to store any data. The system stores web page titles and addresses in the file data module.

When the first index is created, it must be updated periodically. Update the index. These include invalid index removal, changed index updates, and index additions for new pages. When the site data constantly changes, the site and indicator data may change. Some pages have been removed, so the files must be removed from the established database. There are also web pages to be modified, which requires searching the corre-

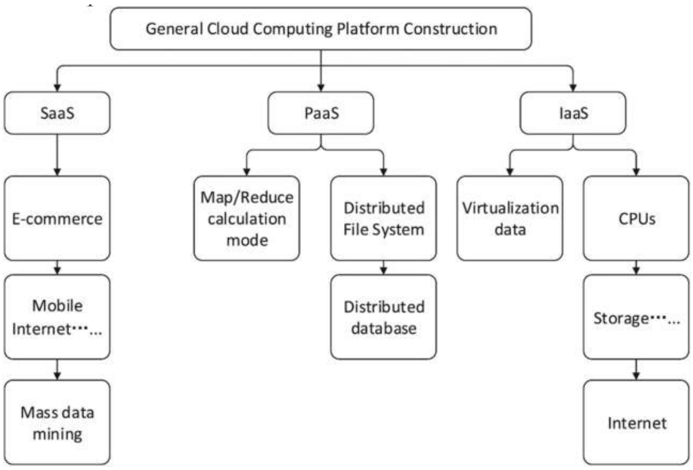

Fig. 2.3: Frame diagram of retrieval submodule.

sponding index library [5]. There are also many newly created web pages. This requires a new index to be created for each new page. Crontab is used to access delta Update under Linux and the site database under Linux. Add, delete, and modify existing databases based on changes in site data. This ensures that the index is consistent with the data on the site.

2.2. Searching for Sub-components. Retrieval submodule is one of the main modules of website search. Conditional queries can be made to the site. Figure 2.3 is a structural diagram of the retrieval subsystem. This article completes an interface similar to standard search engines like Google. Users can type in keywords to search and specify the search criteria [6]. The extraction submodule extracts the keywords entered by the user. Use the segmentation module to divide keywords into several words. According to the user input, the standard Boolean query is constructed. Use collection operations to filter the obtained data. The index database must do further work when a matching set of records is retrieved. This involves deleting duplicate records and classifying records according to time, importance, and other factors. The search results are then returned to the network interface as a list.

2.3. Word segmentation module. Chinese segmentation technology is to divide Chinese into several independent characters. The segmentation module is needed to subdivide the segmented text in the information query, whether in constructing an index or the query [7]. Since Xapian cannot use Chinese automatic segmentation, third-party Chinese automatic segmentation software must be used. There are three kinds of Chinese automatic segmentation methods: automatic segmentation based on a string, automatic segmentation based on understanding and automatic segmentation based on statistics. Currently, Chinese automatic machine-cutting technology mainly includes SCWS, Fudan NLP, ICTCLAS, HTTPCWS, etc. This paper introduces a lexical-based machine Chinese automatic machine segmentation system. Fudan NLP is a software designed by the Java company to support Chinese NLP. ICTCLAS is an open-source Chinese automatic machine independently developed by the Chinese Academy of Sciences [8]. It won the first prize in the National 973 Project evaluation. The splitting speed is about 500 KB/s. The segmentation accuracy reaches 98.45%. HTTPCWS is an open-source automatic Chinese text-cutting system based on HTTP. Its kernel uses ICTCLAS3.02009 to split it.

3. Topology distributed node aggregation method. Firstly, each node's network coordinates are obtained using the distributed network coordinates calculation method based on Vivaldi. Place nodes in a 2D plane (Figure 4). The Euclidean distance can represent the network delay between nodes [9]. The following form of description is taken:

Set the source node to u. ε_i is u concentric circle centered on A. Form cluster $P = \{\varepsilon_i || i = 1, 2, \ldots, n\}$ of concentric circles. Let l_i be the radius of ε_i.

Definition 1. The radii of adjacent concentric circles $\varepsilon_i, \varepsilon_{i-1}$ centered on u are respectively l_i, l_{i-1}. Let's call $l_i - l_{i-1}$ the distance between $\varepsilon_i, \varepsilon_{i-1}$.

Definition 2. Cluster of concentric circles centered on u is cluster of concentric circles centered on $P = \{\varepsilon_i || i = 1, 2, \ldots, n\}$. If $l_i l_{i-1} = l_{i-1} - l_{i-2} = \ldots = l_2 - l_1 = l, l$ is the radius of the inner ring ε_1 of the concentric circle. P is called isometric cluster of concentric circles. P is used to divide the two-dimensional plane of the node [10]. If the innermost circular surface area enclosed by ε_i is and the torus area enclosed between G_1, ε_i and ε_{i-1} is $G_i, i \geq 2$. R_{G_i} is the area of G_i. There is the following lemma:

Lemma 1. $\forall \varepsilon_i \in P, i \geq 2$. Make multiple centripetal lines from the point on ε_i to the center of the circle intersecting ε_{i-1}. The series centripetal lines divide the torus region $G_i, i \geq 2$ into j degrees. So, get $\{G_i = \bigcup g_{i,t} | i = 1, 2, \cdots, n, t = 1, 2, \cdots, j\}$. If $R_{g_{i,t}} = R_{g_{i-1,t}}, i > 2, R_{g_{2,t}} = R_{G_1}$ then $j = 2i - 1$.

Proof: Because$R_{G_i} = \pi(il)^2 - \pi[(i-1)l]^2 = \pi l^2(2i-1)$ and

$$\begin{array}{l} R_{G_i,t} = R_{G_{i-1,t}} = \cdots = R_{G_2,t} = R_{G_c} = \pi l^2, \\ R_{G_i} = jR_{G_{i,t}} = jR_{G_1} = j\pi l^2 \end{array},$$

so $j = 2i - 1$.

You've obtained the certificate. $\forall \varepsilon_i \in P, i \geq 2, \varepsilon_i$ is equally divided by $2i - 1$ arcs. If $\varepsilon_i = \bigcup\limits_{t}^{2i-1} arc_{i,t}, arc_{i,t}$ is an t arc over ε_i, then the length of $arc_{i,t}$ when $i \to \infty$ is πl. l is the inner radius in P.

Proof: First find the length of each arc onε_i after $2i - 1$ equal division

$$\begin{array}{l} c_t, t = 1, 2, \ldots, 2i-1 : c_t = 2\pi \cdot il/2i - 1 \\ \lim\limits_{i \to \infty} c_t = 2\pi l \cdot \lim\limits_{i \to \infty} (i/2i - 1) = \pi l \end{array}$$

proof.

According to Lemma 1, if the torus G_i of P concentric cluster A is to be divided in half so that the area of each subring is equal to the region of the inner ring, then the torus should be divided into odd degrees of equality. The coordinate system of a 2D grid is divided into $1 + 3 + \cdots + 2i - 1 = i^2$ subfields according to the area of equal height. The distance difference l between each subring's outer and inner circle of each subring is found [11]. Each node in the network is positioned according to a particular position under given conditions. Lemma 2 gives that the outer arc length of any subregion converges to πl on the boundary of E. It is also necessary to split the maximum distance between any two points on the subring at $i \to \infty$. There are a couple of theorems here.

Theorem 1. $\forall g_{i,t} \in G_i, i \geq 2$ and $\exists n_1, n_2 \in g_{i,t}, t = 1, 2, \ldots, 2i-1, n_1, n_2$ are points. Define σ as the longest line between n_1, n_2. In this case, $\sigma = l\sqrt{1 + \pi^2}, l$ is the radius of the inner ring in the circle, $i \to \infty$.

4. Performance analysis and experimental simulation. In this part, the TANRA algorithm is verified by a node simulation experiment. Select GT-ITM as the tool for network topology generation. They are using Waxman's random graph model. These parameters are set to alpha and beta of 0.5 and 0.5, respectively. The topological nodes are organized according to the hierarchy of transfer roots, and a delay is added to the edge of the nodes. In total, there are 27 transmission nodes and 436 Stub topology nodes. The boundary delay between the regions is equally spaced along the interval [10,15]. The edge delay between the transmission and Stub areas is equally spaced along the area [40,80]. There is no boundary connection between root domain names. The connection delay between topological nodes in each Stub area is equally spaced along [10,30]. The average degree of each topology node calculated by GT-ITM is 3.55. The number of terminal nodes is gradually increased during the test. The maximum number of terminal nodes can be increased to 10,000.

Zipf assigns the number of terminals occupied by topological nodes. The number of adjacencies of each end node conforms to the Zipf assignment [12]. In each trial, nodes were randomly selected as source nodes. Semantic Routing Improvement algorithm (SCSRAA), DHT algorithm and TANRA algorithm are used to compare the performance of the algorithms. Both SCSRAA and TANRA use Bamboo as a DHT routing algorithm. The DHT algorithm uses a hierarchical search method for multiple adjacent nodes. The number

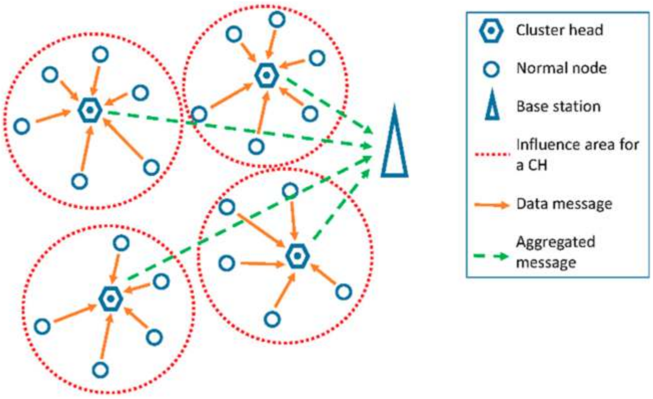

Fig. 3.1: Regional division.

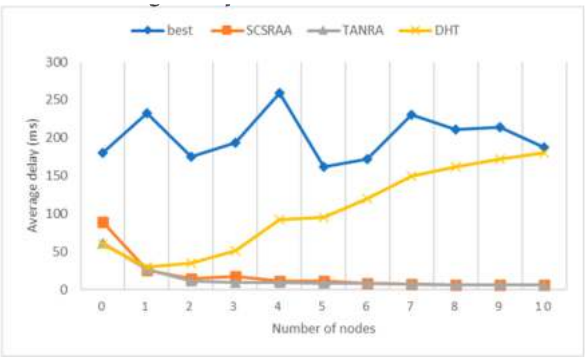

Fig. 4.1: The average latency for the number of adjacent nodes following Zipf (10).

of queries is the same as the previous two methods to ensure openness [13]. Unless otherwise specified, the number of searches is set to 10.

Figure 4.1 shows the average delay in assigning nodes according to Zipf (10) with the number of adjacent nodes at a layer spacing of 30 ms. When the number of access terminal nodes increases, the delay distribution of the SCSRAA system shows prominent fluctuation characteristics. TANRA has good delay convergence. When the number of nodes in the network exceeds 2000, the calculated result is very close to the minimum delay of the network [14]. When the number of networks is less than 1000, the DHT algorithm only exchanges 100 adjacent at a time. Its search algorithm has a significant effect on reducing the average delay. However, when the number of nodes in the network increases, the information interaction between neighbours is insufficient, increasing the network's average delay.

Figure 4.2 shows the average latency of nodes when the number of neighbouring nodes meets the Zipf (40) assignment. With the increase in the number of adjacent nodes, the delay of the TANRA network gradually decreases and tends to the best state [15]. However, SCSRAA delays show significant fluctuations. The DHT algorithm adds 40 neighbours, increasing the system's communication overhead. The average latency increases slightly more slowly than in Figure 5. Still, it continues to rise. When all nodes know only local messages, the network's performance depends on the number of communications between neighbouring nodes and the number of nodes in the entire network. This is consistent with our experimental data.

The introduction of TANRA into the overlay network enables the transformation from 2D to multi-level namespaces. This makes it possible for two adjacent nodes to be split into two sectors corresponding to adjacent segments. This article calls it mismatching [16]. This will affect the success rate of topology matching.

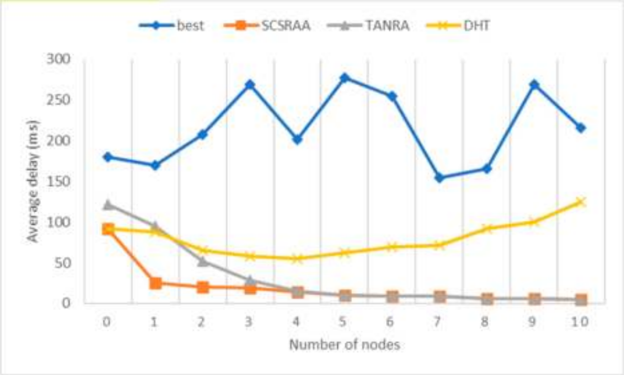

Fig. 4.2: Average latency for the number of adjacent nodes following Zipf (40).

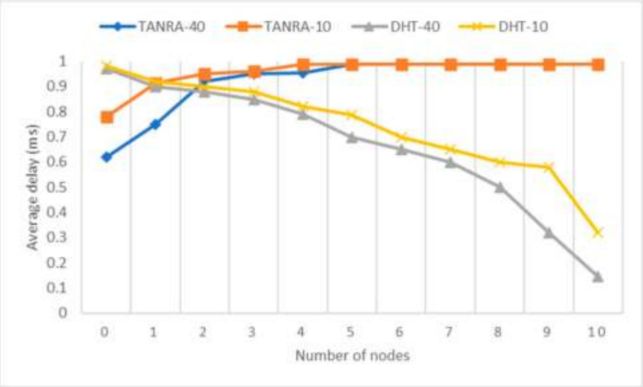

Fig. 4.3: Probability of success for topology comparison.

Figure 4.3 shows the success rate of pairing with a different number of neighbours at a stratified interval of 10 milliseconds. In the case of 1000 nodes, the algorithm's success rate is 65.62% when the neighbour ratio assigned by Zipf (10) is met. At 4000 nodes, the efficiency of the system has reached 98.48%. When the number of nodes in the network meets Zipf (40) allocation and the number of nodes reaches 1000, the transmission success rate of the network can reach 82.29%. In the case of more than 3000 nodes, the success rate of topology comparison of the network is more than 99%. Experimental results show that the TANRA method can effectively improve the topology matching rate of the network when there are a large number of nodes [17]. When only ten adjacent messages interact, the matching rate of the DHT algorithm decreases with the increase in the number of nodes in the network. The reason is insufficient information exchange between adjacent nodes in this method.

Figure 4.4 shows the node has entered the system and 40 neighbour nodes are found in the node and the average latency of the node in the node with a layer spacing of 30 ms. SCSRAA proposes a hierarchical addition method in the case of sparse networks. But when the number of nodes in the network increases, the network moves to a deeper level [18]. This leads to a delay in joining. The TANRA algorithm shows the opposite property. When the network becomes sparse, the nodes have a high probability of no node registration in the region, and their joining delay is relatively high. However, when the number of nodes in the network increases, subsequent nodes join the initial interval with a higher probability. Its additional delay tends to decrease on the whole.

5. Conclusion. In this paper, a distributed node aggregation method, TANRA, based on network coordination and DHT, is designed. The method uses concentric ring clustering with equal spacing to divide the

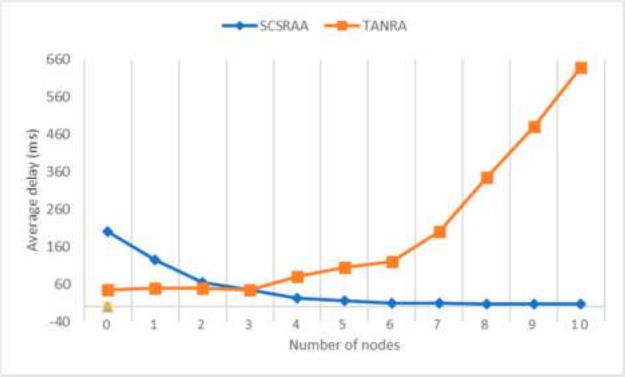

Fig. 4.4: Average delay of increased delay nodes.

space and the area of two nodes on a 2D grid. Then, a local partition method based on multi-level namespaces is proposed. It maintains proximity in each node. In DHT, the two fundamental elements of ”publish” and ”search” are used to add new nodes and find neighbouring nodes. Simulation experiments show the effectiveness of the method. The TANRA method can effectively ensure the consistency of network topology in the case of many nodes while reducing the addition delay. This project will integrate this method with Mesh technology to establish a no-structure overlapping network model to overcome the mismatch problem in the TANRA method. At the same time, the TANRA method is used to optimize the adjacency route in the structured network to reduce the path delay.

Acknowledgments. The authors acknowledge the : This article is a phased research achievement of the 2023 project of Hunan Province’s Education Science ”14th Five Year Plan”: Analysis and Intervention Research on Online Learning Based on Education Big Data Mining (project number: XJK23BXX002).

REFERENCES

[1] Bohu, L., Lin, Z., & Xudong, C. Introduction to cloud manufacturing. Zte Communications, 2020;8(4): 6-9.
[2] Fatemidokht, H., Rafsanjani, M. K., Gupta, B. B., & Hsu, C. H. Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems, 2021; 22(7): 4757-4769.
[3] Murshed, M. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., & Hussain, F. Machine learning at the network edge: A survey. ACM Computing Surveys (CSUR), 2021; 54(8): 1-37.
[4] Mahmood, A., & Wang, J. L. Machine learning for high performance organic solar cells: current scenario and future prospects. Energy & environmental science, 2021; 14(1): 90-105.
[5] Liu, Y., Yuan, X., Xiong, Z., Kang, J., Wang, X., & Niyato, D. Federated learning for 6G communications: Challenges, methods, and future directions. China Communications, 2020; 17(9): 105-118.
[6] Lao, L., Li, Z., Hou, S., Xiao, B., Guo, S., & Yang, Y. A survey of IoT applications in blockchain systems: Architecture, consensus, and traffic modeling. ACM Computing Surveys (CSUR), 2020; 53(1): 1-32.
[7] Zhang, Q., Xin, C., & Wu, H. Privacy-preserving deep learning based on multiparty secure computation: A survey. IEEE Internet of Things Journal, 2021; 8(13): 10412-10429.
[8] Lin, X., Wu, J., Mumtaz, S., Garg, S., Li, J., & Guizani, M. Blockchain-based on-demand computing resource trading in IoV-assisted smart city. IEEE Transactions on Emerging Topics in Computing, 2020; 9(3): 1373-1385.
[9] Ma, Y., Wang, Z., Yang, H., & Yang, L. Artificial intelligence applications in the development of autonomous vehicles: A survey. IEEE/CAA Journal of Automatica Sinica, 2020;7(2): 315-329.
[10] Lindsay, G. W. Convolutional neural networks as a model of the visual system: Past, present, and future. Journal of cognitive neuroscience, 2021; 33(10): 2017-2031.
[11] Ma, L., & Sun, B. Machine learning and AI in marketing–Connecting computing power to human insights. International Journal of Research in Marketing, 2020; 37(3): 481-504.
[12] Lu, H., Zhang, M., Xu, X., Li, Y., & Shen, H. T. Deep fuzzy hashing network for efficient image retrieval. IEEE transactions on fuzzy systems, 2020; 29(1): 166-176.
[13] Jarada, T. N., Rokne, J. G., & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. Journal of cheminformatics, 2020; 12(1): 1-23.

[14] Jiang, J., Chen, M., & Fan, J. A. Deep neural networks for the evaluation and design of photonic devices. Nature Reviews Materials, 2021; 6(8): 679-700.

[15] Gai, K., Guo, J., Zhu, L., & Yu, S. Blockchain meets cloud computing: A survey. IEEE Communications Surveys & Tutorials, 2020; 22(3): 2009-2030.

[16] Chen, H., Jiang, B., Ding, S. X., & Huang, B. Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives. IEEE Transactions on Intelligent Transportation Systems, 2020; 23(3): 1700-1716.

[17] Xin, W. A. N. G., Zi-Yi, W. A. N. G., Zheng, J. H., & Shao, L. I. TCM network pharmacology: a new trend towards combining computational, experimental and clinical approaches. Chinese Journal of Natural Medicines, 2021; 19(1): 1-11.

[18] Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Chen, X., & Wang, X. A comprehensive survey of neural architecture search: Challenges and solutions. ACM Computing Surveys (CSUR), 2021; 54(4): 1-34.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 3, 2023
Accepted: Nov 22, 2023