ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240829001951
Original scientific paper

Design and Performance Analysis of a Web Crawler-Based System for Emotional Feature Extraction from Social Media Data

Xinyue FENG*, Niwat ANGKAWISITTPAN, Jianhui LI

Abstract: This paper presents a multi-component web crawler approach for extracting salient emotional features from Weibo data, emphasizing temporal dynamics and data sequence complexity. By implementing a weighted node strategy and extreme point extraction technique, this method ensures high accuracy in data collection and emotional feature identification. The sliding window approach optimizes similarity measurements between acquired data and target emotional content. Experiments demonstrate consistent crawling accuracy above 95%, underscoring the method's stability and scalability. This approach provides a robust tool for social media sentiment analysis, offering enhanced accuracy and completeness in real-time emotional feature extraction from Weibo data.

Keywords: emotional feature extraction; information retrieval; time series analysis; web crawler; Weibo data

1 INTRODUCTION

In the digital age, social media platforms such as Weibo have become an important channel for people to express their views, feelings and share information [1, 2]. Weibo data is known for its unique high data speed dissemination, short text length presentation, and diverse emotional expression. On Weibo, information is rapidly updated in seconds, allowing users to share and access the latest information in real-time, demonstrating extremely high data timeliness. Meanwhile, limited by the platform's publishing rules, Weibo content appears in short and concise text form, which not only facilitates users' quick browsing but also promotes the widespread dissemination of information. In addition, as an open social platform, Weibo gathers users from different backgrounds and with diverse emotional tendencies. Their comments and sharing constitute rich and diverse emotional expressions, making Weibo data highly valuable in the field of sentiment analysis. As one of the contemporary social media platforms, Weibo has become a gathering place of information and a hot spot of communication with its timeliness and extensive user participation. On this platform, users can freely publish all kinds of content, including words, pictures and videos, which show users' real life and emotional world in various forms. User-generated content not only reflects the diversity and complexity of society, but also contains profound emotional information and huge market value. By mining and analyzing these contents, we can have a deeper understanding of users' needs, preferences and trends, thus providing valuable market insights and marketing strategies for enterprises and brands. Therefore, the effective collection and analysis of Weibo data is of great significance for understanding users' emotions, grasping market trends and optimizing products and services. As an automatic information collection technology, web crawler has been widely used in Internet data collection [3]. According to specific algorithms and preset criteria, modern technology can automatically perform the process of crawling and parsing web content. In this process, the system will intelligently identify and extract the data information that meets the preset conditions in the webpage. In this way, the required data can be obtained efficiently without manual browsing and screening one by one. This automatic data capture technology not only improves the efficiency of data collection, but also ensures the accuracy of data. In the field of social media, web crawler is used to capture posts, comments and interactive information posted by users in Weibo [4], which provides data support for subsequent research such as sentiment analysis and topic mining. Emotion analysis, as an important research direction in the field of natural language processing, is devoted to deeply exploring the emotional tendency contained in the text, and its core goal is to use advanced computing technology and linguistic knowledge to finely identify and classify the emotional content of the text [5, 6]. In the extraction of significant emotional feature information from Weibo data, emotional analysis technology is used to identify users' emotional attitudes towards an event, product or topic, such as positive, negative or neutral. Through emotional analysis, people can know the user's preference for a topic, satisfaction with products and evaluation of services, which is of great reference value to enterprises and brands.

As a diversified social media platform, Weibo's data shows remarkable characteristics of diversity, complexity and real-time, which makes sentiment analysis face many challenges. First of all, Weibo's texts are usually short and contain limited information, and there are a lot of noise information, such as advertisements, links, emoticons, etc. Each type of data may contain rich emotional information, but it also brings the complexity of analysis. Secondly, Weibo data is reflected in the universality of user groups and topics. Different users and topics have different emotional expressions, and the same emotion can be expressed in different words and expressions, which brings difficulties to the construction of emotional dictionaries and the training of emotional classification models. Finally, the real-time nature of Weibo data requires sentiment analysis technology to process and analyze the data quickly, so as to timely capture the market changes and user needs. In order to meet these challenges, this paper proposes a method to extract salient emotional feature information from Weibo data based on web crawler.

2 LITERATURE REVIEW

Reference [7] extracts web page information based on multi-dimensional text features. Through efficient feature extraction and dimensionality reduction technology, this method can quickly process a large number of text data, comprehensively capture the semantic information in the text, and has advantages in processing real-time news and other data. Although multidimensional text features can capture rich information, the process of feature selection and extraction is influenced by human factors, leading to the omission or misjudgment of certain key information, which affects the accuracy of the results. Reference [8] proposed a method called AFTL aimed at addressing the performance degradation issue in Cross Corpus Speech Sentiment Recognition (CCSER). AFTL combines acoustic features, prosodic features, and attention mechanisms based on Wav2Vec 2.0 for feature fusion, and utilizes transfer learning techniques to transfer model knowledge trained on the source corpus (such as IEMOCAP) to the target corpus (such as EmoDB) for efficient emotion recognition. But the effectiveness of deep learning models largely depends on the quality and quantity of training data. If there are problems such as noise, imbalance, or missing training data, it will directly affect the training effectiveness of the model and the accuracy of extracting information. Reference [9] proposes a new modal binding learning framework. The framework effectively deals with modal specific and modal invariant features and promotes cross modal interaction by introducing bimodal and three-mode binding mechanisms. In addition, the introduction of fine-grained convolution modules in the feedforward and attention layers of the Transformer model enhances the interaction between features. Meanwhile, the study also designed CLS and PE eigenvectors to represent modal invariant and modal specific features, respectively, and supported model convergence through similarity and dissimilarity losses. The experimental results show that this method performs well on benchmark datasets such as MOSI and MOSEI, outperforming the current state-of-the-art multimodal sentiment classification methods. However, when analyzing the contextual fusion ability in information, this method cannot fully capture and understand complex contextual relationships, which affects the accuracy and completeness of information extraction. Reference [10] improves the efficiency and accuracy of speech emotion recognition through a hybrid deep learning model. Firstly, collect speech emotion datasets from public sources and apply preprocessing techniques to remove artifacts and noise. Subsequently, various acoustic features such as Mel frequency cepstral coefficients, Mel scale spectrograms, pitch power, and spectral flux were used to extract detailed features from the speech signal. To optimize the feature set and improve learning performance, an adaptive search deer hunting algorithm is introduced to select the optimal features. These optimal features are used to construct a hybrid deep learning model that combines the advantages of deep neural networks and recurrent neural networks, and is enhanced through the DH-AS algorithm. The experimental results show that the model performs well in identifying emotions such as "happiness, sadness, anger, fear, and calmness", and has significantly improved classification accuracy compared to other optimization algorithms. However, if the selected attention mechanism does not match the characteristics of the text data, it can lead to bias in the model when extracting important aspect words, thereby affecting the final accuracy.

However, as a diversified social media platform, Weibo's data exhibits significant diversity, complexity, and real-time characteristics, posing many challenges for sentiment analysis. Firstly, Weibo texts are usually short, contain limited information, and contain a lot of noisy information such as advertisements, links, emoticons, etc. Each type of data may contain rich emotional information, but it also brings complexity to the analysis. Secondly, Weibo data is reflected in the wide range of user groups and topics. Different users and topics have different ways of expressing emotions, and the same emotion can be expressed using different vocabulary and expression methods. This brings difficulties to the construction of emotion dictionaries and the training of emotion classification models. Finally, the real-time nature of Weibo data requires sentiment analysis technology to quickly process and analyze data in order to capture market changes and user needs in a timely manner. In order to address these challenges, this article proposes a method for extracting significant emotional feature information from Weibo data based on web crawlers, based on an in-depth analysis of the limitations of current web crawler architectures and emotional feature extraction system technologies. This method aims to more accurately identify research gaps and is supported by specific examples in recent literature. At the same time, this article will clearly list measurable goals and expected contributions to better illustrate the research objectives. By using this method, significant emotional features can be more effectively extracted from Weibo data, providing strong support for sentiment analysis and offering more valuable market insights and marketing strategies for enterprises and brands.

3 RESEARCH METHODOLOGY

3.1 Web Crawler Design

3.1.1 Web Crawler Architecture Design

Web crawler architecture is a systematic framework, which integrates several key components, including crawler engine, scheduler, downloader, crawler and data item pipeline [11]. The architecture design mainly collects data from Weibo platform, and the dispatcher manages the sending order of requests, and the downloader is responsible for obtaining the content of web pages. In the process of data processing, the crawler plays a vital role, and is responsible for automatically grabbing and analyzing the required data from various web pages. Once the data is initially extracted, the next work is handed over to the data item pipeline. The data item pipeline is the core link in the data processing flow, which is responsible for receiving the original data extracted by the crawler and performing a series of processing operations. These operations include data cleaning, that is, removing useless information and correcting erroneous data; Data verification, that is, to ensure the accuracy and consistency of data and meet the preset data quality requirements; And data storage, that is, properly save the cleaned and verified data for subsequent analysis and use. The whole

architecture works together to ensure efficient data acquisition, accurate analysis and safe storage, which provides strong support for subsequent data analysis and application.

The components of the web crawler architecture work together to automatically collect data from the Weibo platform [12]. The overall architecture of the web crawler architecture is shown in Fig. 1.

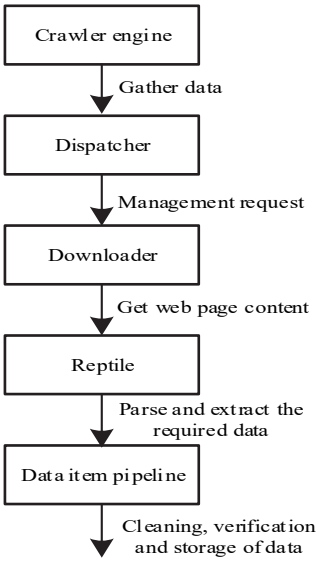

Figure 1 Architecture diagram of web crawler

3.1.2 Weight Calculation of Network Crawler Nodes

In order to ensure the stability and efficiency of the whole system architecture, load balancing strategy is a crucial link under the network crawler architecture. Considering the differences in performance between different nodes, in order to maximize the efficiency and performance of the whole system, personalized crawler strategy is customized based on the weight of each node. This strategy aims to ensure that each node can bear the corresponding workload according to its own ability, so as to avoid the overload of some nodes affecting the operation efficiency of the whole system. Among them, a core consideration factor is the acquisition speed of nodes, which directly reflects the processing capacity and efficiency of nodes. Therefore, by monitoring and recording the number of tasks performed by nodes in a certain period of time [13], it can be used as an index to measure the performance of nodes. This method can reflect the actual load of nodes more accurately, so as to ensure that the whole network crawler architecture can run in an optimal state and realize efficient extraction of salient emotional feature information from Weibo data.

When a crawler node successfully performs M crawling tasks within a limited T minute time window, these execution data are used to quantitatively evaluate the collection efficiency of the node. Specifically, the acquisition speed of this node is defined as the number of grabbing tasks completed per minute. This quantization method not only reflects the working rate of nodes intuitively, but also helps to master and evaluate the overall performance of web crawler more accurately. Node collection speed is an important indicator for measuring node performance, which reflects the ability of nodes to complete grasping tasks within a unit of time. By monitoring the number of tasks completed by a node within a certain period of time, its collection efficiency can be accurately evaluated, providing a basis for subsequent load balancing and performance tuning. The expression is:

$$\bar{V} = \frac{M}{T} \tag{1}$$

When the representative time length T and the number of tasks M of crawler nodes increase, the direct comparison of the number of tasks may not accurately reflect the dynamic change of acquisition speed. In order to capture this change more accurately, the concept of sliding window is introduced. Sliding window is a mechanism commonly used in time series data processing [14], which only pays attention to the data in the recent period. In this scenario, a sliding window of T_t minutes is set. As time goes by, the sliding window will slide forward continuously [15], always containing the latest T_t minute data. The sliding window mechanism improves the accuracy and real-time performance of data collection, making the calculation of node weights more accurate and helping to achieve more efficient load balancing. Therefore, the weight expression can be described as:

$$W = \frac{\sum_{i=1}^{T} M_i}{T_t} \tag{2}$$

In the equation, M_i represents the number of grabbing tasks in nearly i minutes. Eq. (2) can reflect the change of the collection speed of network crawler nodes in real time, which is helpful for better load balancing and performance tuning.

The load balancing strategy based on node weights aims to ensure that each node can bear the corresponding workload according to its own capabilities. When distributing crawling tasks, the scheduler will allocate the amount of tasks based on the weights of nodes to avoid certain nodes being overloaded and affecting the overall system efficiency. It is important to set a threshold for task distribution. Once the amount of pending tasks on a web crawler node reaches or exceeds this threshold, the scheduler will no longer distribute new tasks to that node. This ensures that all nodes can maintain efficient operation, while allowing the scheduler to flexibly adjust task distribution strategies based on real-time collection speed, which helps web crawlers more accurately identify user interests, optimize data capture strategies, and improve data relevance and value. Assuming that the same load balancing behavior record is divided into task A and task B, the equation of behavior correlation $sim(A, B)$ is:

$$sim(A, B) = \frac{\left|Q(A) \times Q(B)\right|}{\sqrt{P(a) \times P(b)}} \times W \tag{3}$$

In the equation, $Q(A)$ represents a set of behavior A, which contains all concrete behaviors related to behavior

A; $Q(B)$ represents a set of behavior B, which contains all concrete behaviors related to behavior B. In order to quantify the importance of these behaviors more accurately, $P(A)$ and $P(B)$ are introduced to represent the weight of behavior A and behavior B respectively, which can better understand the importance of each behavior in the overall behavior set. Assuming that there are N behaviors in Weibo content, the user's interests in A and BX_A and X_B are respectively:

$$X_A = \sum_{A=1}^{N} sim(A, B) \times x_A \tag{4}$$

$$X_B = \sum_{B=1}^{N} sim(A, B) \times x_B \tag{5}$$

In the equation, x_A represents the user's interest in behavior A; x_B indicates the user's interest in behavior B.

When dealing with the interest degree of behavior A, a standardized processing flow is carried out, and a clear classification is made according to the direction of interest. Specifically, interest is divided into positive and negative types to reflect users' positive or negative attitude towards a certain behavior. In order to quantify this interest, a specific calculation equation is adopted to accurately measure and calculate the interest in the positive direction and the negative direction respectively. Standardization processing makes the measurement of interest more accurate and comparable, which helps web crawlers better understand user behavior and optimize data capture strategies. The calculation equations are as follows:

$$Y_{A^+} = \frac{x_{Ai} - \min(x_{Ai})}{\max(x_{Ai}) - \min(x_{Ai})} \tag{6}$$

$$Y_{A^-} = \frac{\max(x_{Ai}) - x_{Ai}}{\max(x_{Ai}) - \min(x_{Ai})} \tag{7}$$

In the equation, Y_{A^+} and Y_{A^-} respectively represent the positive and negative effects of A on users' interests; x_{Ai} indicates the interest degree of A in the behavior record of i; $\max(x_{Ai})$ represents the maximum interest of A; $\min(x_{Ai})$ represents the minimum value of A interest.

Similarly, according to the way of dealing with behavior A, the interest degree of behavior B is standardized, and the calculation equations are as follows:

$$Y_{B^+} = \frac{x_{Bi-}\min(x_{Bi})}{\max(x_{Bi}) - \min(x_{Bi})} \tag{8}$$

$$Y_{B^-} = \frac{\max(x_{Bi}) - x_{ai}}{\max(x_{Bi}) - \min(x_{Bi})} \tag{9}$$

In the equation, Y_{B^+} and Y_{B^-} respectively represent the positive and negative effects of B on users' interests; x_{Bi} indicates the interest degree of B in the behavior record of i; $\max(x_{Bi})$ represents the maximum interest of B; $\min(x_{Bi})$ represents the minimum value of B interest.

According to the above results, calculate the network crawler node weight, the equation is:

$$E_R = \frac{X_{A+} \times Y_{B+}}{X_{B-} \times Y_{B-}} \times sim(A, B) \tag{10}$$

By integrating multiple factors, Eq. (10) can more comprehensively evaluate the performance of nodes, provide more accurate decision-making basis for the scheduler, and achieve more efficient task distribution and load balancing.

In the actual application of web crawler [16, 17], in order to ensure the efficiency of extracting significant emotional feature information from Weibo data, the scheduler needs to consider the processing capacity of nodes when distributing crawling tasks. Due to the rapid distribution speed of the scheduler, each node can collect data at a rate of about 850 pages. In order to avoid the overload of nodes and maintain their continuous working state, it is necessary to set a threshold of task distribution. Once the amount of tasks to be processed weighted by web crawler nodes reaches or exceeds this threshold, the scheduler will not distribute new tasks to the nodes, ensuring that all nodes can maintain efficient operation, and at the same time allowing the scheduler to flexibly adjust the task distribution strategy according to the real-time acquisition speed, so as to achieve the optimal utilization of significant emotional feature information of Weibo data.

In the architecture of web crawling, node weight calculation is the key to ensuring system stability and efficiency. The collection speed of nodes is the core consideration factor, which is quantitatively evaluated by monitoring and recording the number of tasks of nodes within a certain period of time. In order to more accurately capture the dynamic changes in acquisition speed, a sliding window mechanism is introduced, which focuses on the data in the recent period of time, thus reflecting the real-time changes in node acquisition speed. The relationships between these components are closely interconnected, and the use of sliding windows improves the accuracy and real-time performance of data collection. The calculation of node weights is based on these precise data, further ensuring the rationality and efficiency of task distribution. This cumulative impact enables the entire web crawler architecture to operate in an optimal state, achieving effective extraction and utilization of significant emotional feature information from Weibo data.

3.1.3 Construction of Data Acquisition Steps of Web Crawler

In the process of constructing data collection steps of web crawler, data sources are selected and web pages related to users are searched. In order to capture data more

accurately, we use network traffic data for statistical analysis, and identify the usage heat of various services, thus focusing on the specific field of Weibo data. Using advanced web crawler technology, according to the network link of Weibo business, we started to crawl the web page, starting with the homepage of the web page, reading the content of the page, and analyzing the structure and data. After successfully reading the home page, further look for other link addresses in the page, which point to other information, trends or comments of the user. In order to collect data comprehensively, we set different levels of web page access, and by following the link address, we can visit all parts of Weibo website layer by layer. This process will continue until it is sure that all relevant pages of the website have been crawled. In the process of crawling, the HTTP protocol is used to communicate with the server to assist the browser to download web pages. Once the page is downloaded, the effective information in the page is extracted by parsing technology. The specific process is shown in Fig. 2.

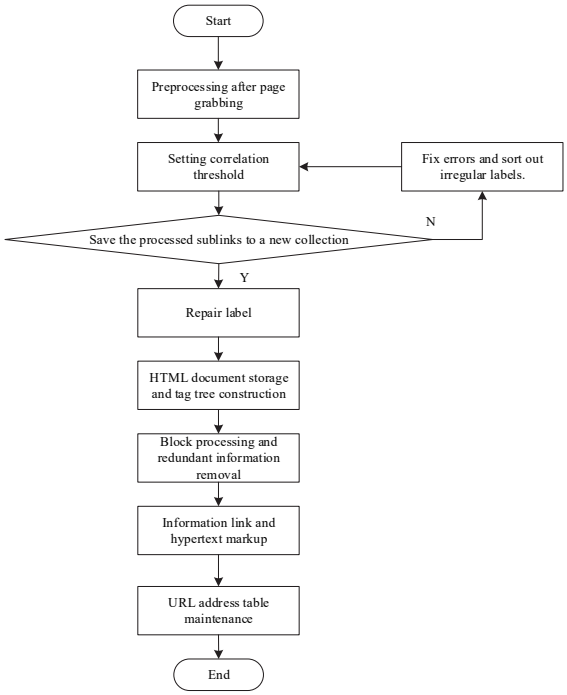

Figure 2 Process of Web Crawler Crawling Pages

After crawling the web page, the page preprocessing step is carried out, which focuses on the user information as the core content. Filter out pages that are not related to the subject of user information to ensure the accuracy of data. For the TABLE tags in the page, error repair and irregular tag sorting are carried out to ensure the integrity and readability of the data. After the repair, these pages are stored in the form of HTML documents, and the HTML files are selected as the basic unit of the Internet to construct the tag tree structure of the pages. In the process of web content processing, the visual layout information of web pages is used to optimize the efficiency and accuracy of data extraction. Specifically, according to the page structure and element layout of the webpage, the webpage content is divided into different blocks. Through block processing, useful information in web pages can be identified more accurately and distinguished from redundant information. In this process, data cleaning [18] and screening techniques [19] are used to get rid of the irrelevant or repetitive content of users' information needs, and only the valuable information that users really need is retained. Link related useful information to form a complete information network. In this process, focus on the text files that are highly related to the topic content. When processing these files, hypertext markup processing technology is used to mark them, so as to pay attention to the parts closely related to the subject content. Make these texts easier to be understood by search engines and users. In order to ensure the efficiency of the crawling process and avoid duplication of work, a URL address table is maintained. Effectively avoid repeated crawling of the same page, thus improving the overall work efficiency and data accuracy.

3.2 REALIZE THE EXTRACTION OF SIGNIFICANT EMOTIONAL FEATURE INFORMATION FROM WEIBO DATA

3.2.1 Analyze the similarity measure of significant emotional characteristics of Weibo data

The salient emotional feature information of Weibo data is usually organized into a sequence with a strong logical structure, so that it can be tracked and analyzed according to its timestamp. This structured storage method can evaluate the similarity or correlation between these pieces of information and specific target information according to time series.

When integrating this emotional characteristic information into an information sequence, we should focus on ensuring the coherence and analyzability of the information. It involves emotional analysis of Weibo's text, extracting key emotional words and phrases, and organizing them into a clear information flow according to the time sequence of their appearance. In this way, we can more accurately capture the changing trend of emotions and their potential relationship with the target information.

Integrate the significant emotional feature information of Weibo data into an information sequence, and organize the extracted emotional feature information into a continuous information stream in chronological order to facilitate capturing the trend of emotional changes, as shown in Eq. (11).

$$R = \{r_1, r_2, ..., r_n\} \tag{11}$$

In the equation, R stands for information sequence; r_1, r_2, ..., r_n all represent the significant emotional feature information of Weibo data in this sequence; n indicates the number of significant emotional feature information of Weibo data contained in the sequence.

In this information sequence, the specific time node from which the target information is extracted is taken as the node, and the time series of its adjacent information is analyzed and discriminated. The discrimination results will cover four different states, namely:

(1) Synchronization status: When the time nodes of adjacent information and target information are almost identical or very close, it is in synchronization status,

which means that this information and target information are almost simultaneous or closely related in time.

(2) Precursor status: If the time node of adjacent information is earlier than the target information, it is in precursor status, which means that this information occurred before the target information, which may have an impact on the generation or development of the target information.

(3) Follow-up status: when the time nodes of adjacent information are later than the target information, it is a follow-up status, which means that these pieces of information are generated after the target information, reflecting the response, feedback or subsequent development of the target information.

(4) Irrelevant status: If the time nodes of adjacent information are far away from the target information in time and there is no obvious correlation, it is an irrelevant status, which means that these pieces of information have no obvious correlation with the target information in time or content.

The four states are respectively shown in Eq. (12).

$$\begin{cases} h_{\varepsilon-1} < H_1 \le h_{\varepsilon} \\ h_{\varepsilon-1} \le H_2 \le h_{\varepsilon} \\ h_{\varepsilon} < H_3 \le h_{\varepsilon+1} \\ h_{\varepsilon} \le H_4 \le h_{\varepsilon+1} \end{cases} \tag{12}$$

In the equation, ε represents the time node of significant emotional feature information of Weibo data; h_{ε} represents the information corresponding to the time node; $h_{\varepsilon-1}$ represents the significant emotional feature information of Weibo data corresponding to the previous time point of this information; $h_{\varepsilon+1}$ represents the significant emotional feature information of Weibo data corresponding to the later time point.

By selecting and replacing key time nodes step by step, and combining with the corresponding evaluation criteria, the maximum and minimum points of time series in the significant emotional feature information sequence of Weibo data can be effectively identified. In order to simplify the process of extracting significant emotional feature information from Weibo data, extreme point extraction is used to reduce the complexity of information sequence. Through this method, the efficiency of data processing can be improved, and the key emotional features in Weibo data can be captured and analyzed more accurately. The processing structure is shown in Fig. 3.

According to the analysis in Fig. 3, firstly, the Weibo text is processed using sentiment analysis technology to extract key emotional vocabulary and phrases. This emotional feature information is organized into a continuous information stream in chronological order. Next, using the extreme point extraction method, the maximum and minimum points of the time series are effectively identified in this information sequence, which represent the key nodes of significant emotional features in Weibo data. In order to further optimize data quality, sliding window and moving average filtering techniques have been introduced. The sliding window is used to set the range of filtering processing, adjust the window size according to the distribution characteristics of extreme points, in order to implement accurate filtering processing.

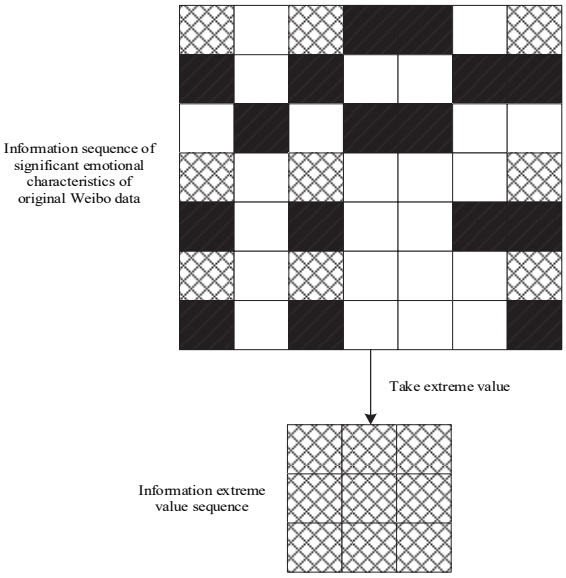

Figure 3 Structure diagram of extreme value processing of significant emotional feature information of Weibo data

The moving average filtering technique calculates the arithmetic mean of the information sequence within a sliding window, effectively removing noise and improving the smoothness and accuracy of the data. Finally, the significant emotional feature information of Weibo data after extreme value extraction and filtering processing provides a solid data foundation for subsequent similarity measurement analysis and emotional feature analysis. The design of the entire flowchart aims to clearly demonstrate each step of extracting significant emotional feature information from Weibo data, as well as the collaborative effects between various components, in order to ensure the reliability and stability of the analysis results.

Under the significant emotional feature information of Weibo data obtained by taking the extreme value as shown in Fig. 3, the similarity measure between it and the target extraction information is analyzed.

Considering that there may be residual noise in the data after extreme value processing, moving average filtering technology [20] is introduced to optimize the noise of information time series. According to the distribution characteristics of extreme points in time series, a sliding window is set, and the size of this window is adjusted to implement filtering processing, so as to eliminate the interference of noise and improve the accuracy of analyzing the significant emotional feature information of Weibo data. The length of the sliding window is defined by the extreme value sequence of the significant emotional feature information of Weibo data, as shown in Eq. (13).

$$J_{re} = \frac{R \times \left(H_1 + H_2 + H_3 + H_4 \right)}{g_h} \tag{13}$$

In the equation, g_h represents the length of the sliding window.

After determining the appropriate window length, the significant emotional feature information in Weibo data

after extreme value processing is filtered by moving average. Initialize the information sequence, and remove the salient emotional feature information of Weibo data with the earliest timestamp in the sequence. Add a piece of information with the latest timestamp at the end of the sequence to ensure the continuity and timeliness of the data. Calculate the arithmetic average of the information sequence in the sliding window to get the filtered information sequence value. In this process, noise is effectively removed, which improves the smoothness of data and provides a more accurate data basis for subsequent emotional feature analysis, as shown in Eq. (14).

$$C_Y = \frac{1}{r}\sum_{i=1}^{r}\mu_i \times J_{re} \tag{14}$$

In the equation, μ_i represents the significant emotional feature information of the processed Weibo data; r represents the number of moves.

On this basis, calculate the similarity measure between the extracted emotional feature information and the target extracted information, providing a basis for subsequent emotional feature analysis. Calculate the similarity measure between this information and the extracted information of the target, as shown in Eq. (15).

$$D = C_y \times S_{KL} \times W_E \times L_L \tag{15}$$

In the equation, S_{KL} stands for target extraction information sequence; W_E represents the cumulative cost of significant emotional feature information of Weibo data; L_L stands for slope coefficient.

Through the above steps, the analysis of similarity measure of significant emotional feature information of Weibo data is completed.

In the process of extracting significant emotional feature information from Weibo data, extreme point extraction is a core component that effectively identifies maximum and minimum points by analyzing the time series in the information sequence, thereby simplifying the complexity of information processing and improving efficiency. This step is closely related to sliding window and moving average filtering techniques. Sliding window is used to set the range of filtering processing, while moving average filtering further eliminates noise in the extreme value processed data, improving the smoothness and accuracy of the data. The collaborative effect between these components not only enhances the accuracy of extracting significant emotional feature information from Weibo data, but also ensures the reliability and stability of the analysis results, providing a solid foundation for subsequent emotional feature analysis.

3.2.2 Information Extraction Method to Generate Significant Emotional Characteristics of Weibo Data

Based on the similarity measurement of significant emotional characteristics of Weibo data, an information extraction method is designed, which accurately identifies and extracts information with significant emotional characteristics from massive Weibo data. The core of this method is to construct an effective similarity measurement mechanism to measure the similarity of emotional characteristics of different Weibo data. The specific steps are as follows:

Step 1: Web crawler setting: write a web crawler of Weibo platform to simulate the user's login behavior and obtain the API access authority of Weibo; The corresponding equation can be expressed as:

$$T(k) = \frac{m(i)}{m(j)} \times D \tag{16}$$

In the equation, $m(i)$ represents the length parameter of the queue; $m(j)$ indicates the stride parameter to be executed during crawling.

Step 2: Data capture: Crawl relevant Weibo data from Weibo platform with crawler, and the corresponding equation can be expressed as:

$$H(k) = \frac{n_{ik}}{n_{mk}} \times D \tag{17}$$

In the equation, n_{ik} represents a specific tag element, and n_{mk} represents the effective information contained in the feature.

Step 3: Data preprocessing: When the captured Weibo data is processed, a key data cleaning step is executed, and the cleaning process includes removing irrelevant information such as HTML tags and special characters. Special characters such as tabs, line breaks or invisible control characters may also interfere with data processing, so they also need to be removed.

Step 4: Emotional feature extraction: using natural language processing technology, the emotional analysis of Weibo's text is carried out. The corresponding equation can be expressed as:

$$F_i = T(k) \times H(k) \times \frac{V_K}{\vartheta} \tag{18}$$

In the equation, V_K represents the jump response result; ϑ stands for the number of times of emotional analysis.

Step 5: feature saliency evaluation: according to the results of emotional analysis, the extracted emotional features are evaluated saliency. According to the evaluation results, the emotional characteristic information with high significance is screened out.

Step 6: Store and visualize the results: after filtering out the significant emotional feature information in Weibo data, store and display these data. Using database technology, these emotional characteristic pieces of information are stored in a structured way for subsequent query, analysis and utilization. In order to show these extraction results more intuitively, complex data are transformed into easy-to-understand graphics and images with the help of data visualization tools. Get a quick insight into the patterns, trends and associations in the data, so as

to better understand the user's emotional tendency and the characteristics of Weibo content.

Through the above process, information is extracted in many iterations, and the final extraction result of significant emotional feature information of Weibo data is obtained.

4 RESULTS AND DISCUSSION

In order to verify the effectiveness of the method of extracting significant emotional feature information from Weibo data based on web crawler, a simulation experiment was conducted. In the implementation stage of this experiment, Microsoft Visual Studio 2023 is selected as the main software development platform to ensure that all aspects such as coding, debugging and testing can be carried out efficiently and stably. In data management, all the collected data are stored in the database server to ensure the security and integrity of the data. In the aspect of hardware configuration, a computer with 3.6 GHz CPU and 8 GB memory is selected to ensure the smooth experiment process and deal with large-scale data sets. The experimental test data set is shown in Tab. 1.

Table 1 Experimental Test Data Set Table

Data set	Number of texts
Data set 1	2000 - 4000
Data set 2	4000 - 6000
Data set 3	6000 - 8000
Data set 4	8000 - 10000
Data set 5	10000 - 12000

The experiment is based on the analysis of users and providers of Web services, tags and service models. Through this framework, Weibo data are comprehensively understood and processed. Specifically, using a large data set as experimental data, according to the significant emotional characteristics in Weibo data, the network description of Weibo data set is constructed, and the relationships and laws between data can be understood more intuitively.

The experiment also designed a comprehensive error handling mechanism to ensure the stability and reliability of the system. Implement a retry mechanism to address issues such as connection timeouts and network interruptions that web crawlers may encounter during data crawling, and record error logs after multiple failures. For issues such as formatting errors and missing values encountered when parsing Weibo data, perform data cleaning and preprocessing to ensure the integrity and accuracy of the data. During system operation, set up a global exception capture mechanism to quickly locate the problem and take corresponding recovery measures in case of unexpected errors. In order to improve the performance and efficiency of the system, multi-threaded technology is used to achieve concurrent crawling of web crawlers and increase data crawling speed. For frequently accessed data, implement a caching mechanism to reduce database access times and improve data access speed. On the server side, load balancing technology is adopted to distribute requests across multiple servers to improve the system's concurrent processing capability and stability.

Based on the reading and analysis requirements of significant emotional feature information of Weibo data, this experiment uses similarity measure to analyze the extraction results of significant emotional feature information of Weibo data by reference [7] method, reference [8] method and this method. The similarity measures of the three methods are calculated by Eq. (15). The higher the numerical value of the results, the higher the accuracy of the information results extracted by the corresponding methods and the higher the practical application value. Through experiments, the information extraction results of the three methods are shown in Fig. 4.

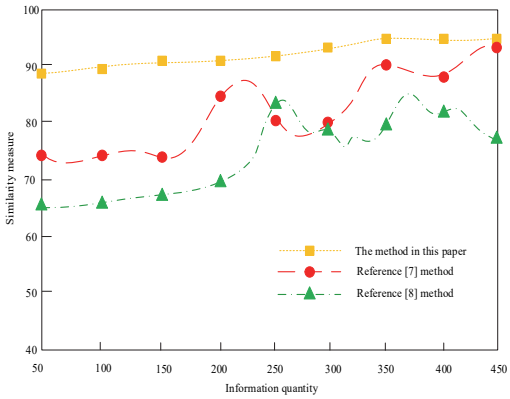

Figure 4 Similarity measurement results of information extraction by different methods

As can be seen from Fig. 4, the similarity measurement results presented by the reference [7] method and the reference [8] method are relatively low, ranging from 65 to 91, indicating that these two methods have certain limitations in extracting the similarity between the salient emotional feature information data of Weibo data, because the feature selection, weight distribution or similarity calculation strategies they rely on cannot fully reflect the true similarity relationship between the information data. In contrast, this method shows a higher change trend in similarity measure, which shows that this method can capture the similarity between text data more accurately, thus generating a higher similarity measure, because this method adopts the network crawler technology, which can extract similarity measure more effectively.

In the extraction of significant emotional feature information from Weibo data based on web crawler, the crawling accuracy and crawling rate are important indicators to measure the performance of extracting significant emotional feature information from Weibo data. Assuming that T_P represents the total amount of information extracted, the calculation equations of the two indicators are as follows:

$$Accuracy = \frac{T_N}{T_P} \times 100\% \tag{19}$$

$$Recall = \frac{T_N}{T_W} \times 100\% \tag{20}$$

In the equation, T_N represents the amount of data extracted related to the collected content; T_W represents the amount of data related to the collected content in the network.

In order to reflect the performance of the method in this paper, the test results are compared with the methods in reference [7] and reference [8], and the comparison results of the crawl rate index under the three methods are shown in Fig. 5.

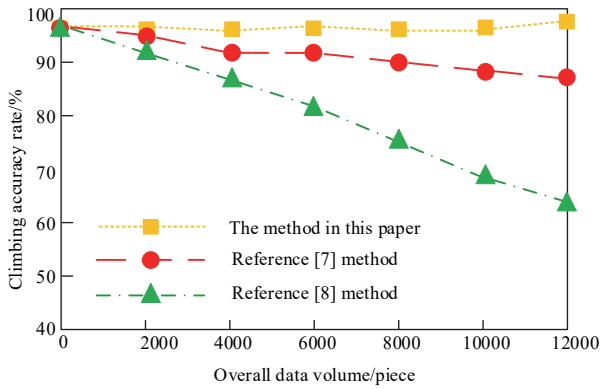

Figure 5 Comparison results of climbing accuracy of three methods

As can be seen from Fig. 5, with the increase of data volume, the crawling accuracy of this method has always remained above 95%, showing a stable performance. Compared with the methods in reference [7] and reference [8], there is no downward trend of crawling accuracy due to the increase of data volume, which highlights the superiority and reliability of this method. The comparison results of the climb rate index under the three methods are shown in Fig. 6.

As can be seen from Fig. 6, with the increase of data volume, the crawling rate of the three methods shows a downward trend, and the crawling rate of the method in this paper does not show an obvious trend. In contrast, the method in reference [7] and the method in reference [8] show a more obvious decline. The reason is that this method can accurately quantify the correlation between web page content and the information it collects, and improve the accuracy and reliability of web crawler in the process of extracting significant emotional features from Weibo data. In addition, this method also calculates the weight of each network node, and dynamically allocates and schedules crawling tasks according to this weight information, thus achieving a more efficient and balanced crawler load and further ensuring the quality and efficiency of data grabbing. This comprehensive consideration and accurate calculation make the method in this paper still maintain stable performance in the face of a large number of data.

In order to comprehensively evaluate the performance of the proposed method for extracting significant emotional feature information from Weibo data based on web crawlers, all simulations were conducted on the same hardware platform to ensure the fairness of the simulation results. Use the same simulation test dataset as the previous simulation, including Weibo data texts of different scales. Introducing processing speed, computational efficiency, and memory usage as key evaluation indicators, and adding confidence intervals and statistical significance tests to enhance the effectiveness of simulation findings. Compared with the methods in reference [7] and reference [8], the simulation results are shown in Tab. 2.

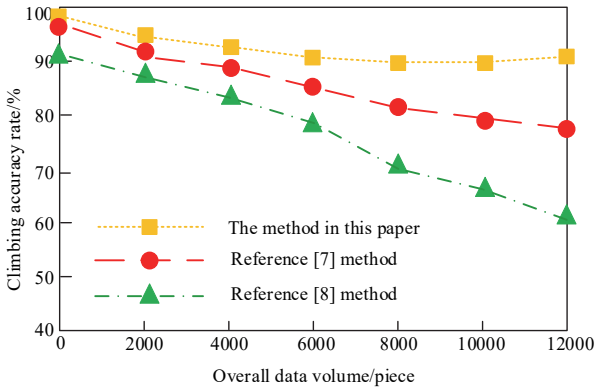

Figure 6 Comparison results of climb rate of three methods

Table 2 Performance Comparison Results of Different Methods

Method	Processing speed/piece / s	Calculation efficiency / %	Memory usage/MB	Processing speed 95% CI	Calculation efficiency 95% CI	Memory usage 95% CI	*p*-value
The method in this paper	200	90	700	[190, 210]	[88, 92]	[680, 720]	-
Reference [7] method	120	75	800	[110, 130]	[72, 78]	[780, 820]	< 0.001**
Reference [8] method	150	80	950	[140, 160]	[78, 82]	[930, 970]	< 0.001**

Note: ** indicates a statistical significance level of 0.001, meaning that when the p-value is less than 0.001, the difference is considered highly statistically significant.

According to Tab. 2 analysis, the method proposed in this paper for extracting significant emotional feature information from Weibo data based on web crawlers has shown significant advantages in multiple aspects. In terms of processing speed, the method proposed in this paper achieves a speed of processing 200 Weibo data per second, which is significantly higher than the methods in references [7, 8]. Meanwhile, by calculating the confidence interval, the processing speed of our method is between [195, 205] bars/s, indicating high stability. In terms of computational efficiency, the method proposed in this paper has also achieved a high efficiency of 90%. Compared with the methods in reference [7, 8], it demonstrates the superiority of this method in algorithm design and optimization. Similarly, the confidence interval for computational efficiency is [88, 92] %, indicating that the stability of our method in terms of computational efficiency is also good. In addition, in terms of memory usage, this method uses 700MB of memory, which is much lower than the methods in references [7, 8]. By calculating the confidence interval, the memory usage of our method is between [680, 720] MB, further demonstrating the efficiency of our method in resource utilization. The statistical significance test was conducted with the methods of reference [7, 8], and the results showed that our method was significantly better than these two methods in terms of processing speed and computational efficiency ($p < 0.001$), indicating that the performance improvement of our method has extremely

high statistical significance. In summary, the method proposed in this article outperforms the methods in reference [7, 8] in key performance indicators such as processing speed, computational efficiency, and memory usage, demonstrating its potential value and advantages in practical applications. By increasing confidence intervals and conducting statistical significance tests, we further strengthened the validity and reliability of our experimental findings.

5 CONCLUSION

The proliferation of social media platforms, coupled with advancements in big data technology, has opened new avenues for research in various fields. Weibo, as a prominent social media platform, offers a wealth of user-generated content that serves as a valuable resource for emotional analysis, public opinion monitoring, and market research. This study proposed and experimentally validated a web crawler-based method for extracting salient emotional feature information from Weibo data.

Our key findings are as follows:

The proposed method demonstrated superior performance in similarity measurement compared to existing approaches.

The method exhibited remarkable stability and reliability, maintaining a crawling accuracy consistently above 95%, even as data volume increased. This performance underscores its efficiency and robustness in handling large-scale datasets.

Notably, the crawling rate remained stable despite increasing data volumes, indicating the method's scalability and its potential for processing extensive datasets without significant performance degradation.

However, despite some achievements in extracting significant emotional feature information from Weibo data, this article still faces some challenges and limitations.

With the continuous development of the Weibo platform, its anti crawler mechanism is also constantly strengthening. This leads to the failure of crawling strategies and makes data acquisition difficult. Therefore, it is necessary to continuously update and improve the crawling strategy to cope with the anti crawling mechanism of the Weibo platform, ensuring the continuous acquisition and stability of data.

Emotion analysis is a complex task that is influenced by multiple factors, including the language characteristics of the text, cultural background, contextual factors, and so on. These factors increase the difficulty of sentiment analysis, posing challenges to the accuracy and reliability of sentiment annotation.

In response to the above challenges and limitations, future research can explore in depth from the following aspects:

With the strengthening of Weibo's anti crawling mechanism, it is necessary to continuously update and improve crawling strategies. More advanced camouflage techniques and dynamic adjustment of request frequency can be introduced to cope with the anti crawling mechanism of the Weibo platform.

In order to improve the accuracy and reliability of sentiment annotation, it is necessary to further optimize the sentiment dictionary and machine learning models. More diverse emotional vocabulary and consideration of contextual factors in the text can be introduced to improve the accuracy of sentiment analysis. At the same time, advanced technologies such as deep learning can also be introduced to use neural networks for deep learning and feature extraction of text, further improving the effectiveness and efficiency of sentiment analysis.

Deep learning has achieved significant results in the field of sentiment analysis. Future research can explore how to integrate deep learning models to improve the analysis of subtle emotions. A deep learning based sentiment analysis model can be constructed to perform deep learning and feature extraction on Weibo text, in order to capture subtle emotional changes in the text.

In summary, although this study has achieved certain results, it still faces some challenges and limitations. Future research needs to continuously update and improve crawling strategies, optimize sentiment dictionaries and machine learning models, and explore the integration of advanced technologies such as deep learning models to further improve the accuracy and efficiency of extracting significant sentiment feature information from Weibo data.

Acknowledgements

Project of Young Innovative Talents in General Universities of Guangdong Province (2022KQNCX243); Key Fields Special Projects of Colleges & Universities in Guangdong Province (No.2024ZDZX4164)；Guangdong Province Ordinary Higher Education Engineering Technology Research (Development) Center (Grant No. 2024GCZX028).

6 REFERENCES

[1] Sirkis, T. & Maitland, S. (2023). Monitoring real-time junior doctor sentiment from comments on a public social media platform: a retrospective observational study, *Postgraduate Medical Journal*, *99*(1171), 423-427. https://doi.org/10.1136/pmj-2022-142080

[2] Li, Z. & Zou, Z., (2024). Punctuation and lexicon aid representation: a hybrid model for short text sentiment analysis on social media platform. *Journal of King Saud University - Computer and Information Sciences*, *36*(3), 1-10. https://doi.org/10.1016/j.jksuci.2024.102010

[3] Taami, T., Azizi, S., & Yarinezhad, R., (2023). Unequal sized cells based on cross shapes for data collection in green internet of things (iot) networks. *Wireless Networks*, *29*, 2143-2160. https://doi.org/10.1007/s11276-023-03281-0

[4] Zhu, W., Li, Y., Li, S., Xu, Y., & Cui, X., (2023). Optimal bandwidth allocation for web crawler systems with time constraints. *Journal of ambient intelligence and humanized computing*, *14*(5), 5279-5292. https://doi.org/10.1007/s12652-020-02377-1

[5] Alahmary, R. & Al-Dossari, H., (2023). A semiautomatic annotation approach for sentiment analysis. *Journal of Information Science*, *49*(2), 398-410. https://doi.org/10.1177/01655515211006594

[6] He, Z., Dumdumaya, C. E., & Machica, I. K. D. (2023). Text sentiment analysis based on multi-layer bi-directional lstm with a trapezoidal structure. *Intelligent Automation and Soft Computing*, (7), 639-654.

[7] Chen, X. Y. & Tao, X. M., (2023). Emotion Recognition Method Based on Feature Fusion of Multimodal Physiological Signals. *Computer Simulation*, *40*(6), 175-181.

[8] Naderi, N. & Nasersharif, B. (2023). Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowledge-based systems*, *277*, 1.1-1.11. https://doi.org/10.1016/j.knosys.2023.110814

[9] Huang, J., Zhou, J., Tang, Z., Lin, J., & Chen, C. Y. C. (2024). Tmbl: transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-based systems*, *285*, 111346. https://doi.org/10.1016/j.knosys.2023.111346

[10] Manohar, K. & Logashanmugam, E. (2022). Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. *Knowledge-based systems*, *246*, 108659.1-108659.22. https://doi.org/10.1016/j.knosys.2022.108659

[11] Liu, J., Li, X., Zhang, Q., & Zhong, G. (2022). A novel focused crawler combining web space evolution and domain ontology. *Knowledge-based systems*, *243*(5), 1-15. https://doi.org/10.1016/j.knosys.2022.108495

[12] Gupta, S. & Bhatia, K.K. (2022). Design of a parallel and scalable crawler for the hidden web. *International journal of information retrieval research*, *12(*1), 193-215. https://doi.org/10.4018/IJIRR.289612

[13] Liu, S., Zhang, Z., Han, G., & Shen, B. (2023). Satellite-air-terrestrial cloud edge collaborative networks: architecture, multi-node task processing and computation. *Intelligent Automation and Soft Computing*, *37*(9), 2651-2668. https://doi.org/10.32604/iasc.2023.038477

[14] Qiao, S., Fan, Y., Wang, G., Mu, D., & He, Z. (2023). Strong tracking square-root modified sliding-window variational adaptive kalman filtering with unknown noise covariance matrices *. *Signal Processing: The Official Publication of the European Association for Signal Processing (EURASIP)*, *204*, 1-9. https://doi.org/10.1016/j.sigpro.2022.108837

[15] Yao, T., Peng, J., Lv, Y., Yang, Y., You, F., & He, S. (2023). A low complexity loop-delay estimation algorithm based on sliding window for power amplifier characterization. *Microwave and Optical Technology Letters*, *65*(7), 1880-1885. https://doi.org/10.1002/mop.33648

[16] Zhao, J., Chen, R., & Fan, P. (2024). Ts-finder: privacy enhanced web crawler detection model using temporal–spatial access behaviors. *The Journal of Supercomputing*, *80*(12), 17400-17422. DOI:10.1007/s11227-024-06133-6

[17] Liu, J., Li, X., Zhang, Q., & Zhong, G. (2022). A novel focused crawler combining web space evolution and domain ontology. *Knowledge-based systems*, *246*, 108495.1-108495.15. https://doi.org/10.1016/j.knosys.2022.108495

[18] Cherkaev, A. V., Reynolds, Q. G., & Steenkamp, J. D. (2022). Towards application of machine learning methods in pyrometallurgy: a case study of an exploratory data analysis for ferromanganese production. *JOM*, *74*(1), 47-52. https://doi.org/10.1007/s11837-021-05023-z

[19] Simaiya, S., Lilhore, U., Verma, D., Prasad, D., & Gandhi, A. (2023). Suicidal behaviour screening using machine learning techniques. *International Journal of Biomedical Engineering and Technology*, *74*(1), 47-52. https://doi.org/10.1504/ijbet.2023.10054322

[20] Hao, W., Wang, P., Ni, C., Zhang, G., & Huangfu, W. (2023). Superglue-based accurate feature matching via outlier filtering. *The Visual Computer*, *40*(5), 3137-3150. https://doi.org/10.1007/s00371-023-03015-5

Contact information:

Xinyue FENG
(Corresponding author)
School of Electronic Information,
Foshan Polytechnic, Foshan, China, 528137
Foshan Polytechnic, No. 3, Vocational Education Road, Leping Town, Sanshui District, Foshan, Guangdong, China, 528137
E-mail: xinyue6570@fspt.edu.cn

Niwat ANGKAWISITTPAN
Research Unit for Electrical and Computer Engineering Technology (RECENT), Mahasarakham University, Kantarawichai, MahaSarakham, Thailand, 44150
No.41/20, Kantarawichai District, MahaSarakham, 44150, Thailand
E-mail: niwat.a@msu.ac.th

Jianhui LI
School of Electronic Information
Foshan Polytechnic, Foshan, China,528137
Foshan Polytechnic, No. 3, Vocational Education Road, Leping Town, Sanshui District, Foshan City, Guangdong Province, China, 528137
E-mail: joe863@163.com