Hindawi
Security and Communication Networks
Volume 2023, Article ID 9873964, 1 page
https://doi.org/10.1155/2023/9873964

Retraction

Retracted: An Algorithm of Scene Information Collection in General Football Matches Based on Web Documents

Security and Communication Networks

Received 26 December 2023; Accepted 26 December 2023; Published 29 December 2023

Copyright © 2023 Security and Communication Networks. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] B. Li and T. Zhang, "An Algorithm of Scene Information Collection in General Football Matches Based on Web Documents," *Security and Communication Networks*, vol. 2021, Article ID 5801631, 11 pages, 2021.

Hindawi
Security and Communication Networks
Volume 2021, Article ID 5801631, 11 pages
https://doi.org/10.1155/2021/5801631

WILEY | Hindawi

Research Article

An Algorithm of Scene Information Collection in General Football Matches Based on Web Documents

Bin Li[1] **and Ting Zhang**[2]

[1]*School of Physical Education and Sport Science, Fujian Normal University, Fuzhou, Fujian Province 350017, China*
[2]*School of Chinese Language, Media, and Law, Fujian Polytechnic Normal University, Fuqing, Fujian Province 350300, China*

Correspondence should be addressed to Ting Zhang; zting3115@163.com

Received 11 June 2021; Revised 14 September 2021; Accepted 17 September 2021; Published 14 October 2021

Academic Editor: Chi-Hua Chen

Copyright © 2021 Bin Li and Ting Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to obtain the scene information of the ordinary football game more comprehensively, an algorithm of collecting the scene information of the ordinary football game based on web documents is proposed. The commonly used T-graph web crawler model is used to collect the sample nodes of a specific topic in the football game scene information and then collect the edge document information of the football game scene information topic after the crawling stage of the web crawler. Using the feature item extraction algorithm of semantic analysis, according to the similarity of the feature items, the feature items of the football game scene information are extracted to form a web document. By constructing a complex network and introducing the local contribution and overlap coefficient of the community discovery feature selection algorithm, the features of the web document are selected to realize the collection of football game scene information. Experimental results show that the algorithm has high topic collection capabilities and low computational cost, the average accuracy of equilibrium is always around 98%, and it has strong quantification capabilities for web crawlers and communities.

RETRACTED

1. Introduction

With the continuous development of football and modern science and technology, sports science and technology workers have carried out some statistics, analysis, and evaluation in sports. The primary task of scene information collection in general football matches is to collect information from various channels. Due to different research contents and purposes, there are obvious differences in scene information in football matches [1]. The real-time football matches and the need for professionals for real-time information require the on-the-spot information collection algorithm of football to be universal, easy to operate, and in real time. However, as a science of understanding and using information, information science provides a new way of thinking and method for developing scene information collection on the football match. Meanwhile, intelligent computing research aims at bringing intelligence, reasoning, perception, information gathering, and analysis to computer systems [2–4]. It provides a new way of thinking for the scene information collection on the football match. Information method is a research method to achieve its purpose by using the acquisition, transmission, processing, and processing of information [5].

With the rapid development of the Internet, the network is profoundly changing our lives. WWW (World Wide Web), the most rapidly developing technology on the Internet, has gradually become the most important way of information release and transmission on the Internet with its intuitive, convenient use and rich expression ability [6]. With the advent and development of the information age, the information on the web is growing rapidly. As of January 2015, the number of web pages on the Internet has exceeded 2.1 billion. The number of Internet users has exceeded 300 million, and the number of web pages is still increasing at 7 million per day. This provides rich resources for people's life. However, the rapid expansion of web information, while providing people with rich football match information, makes people face a huge challenge in the effective use [7]. On the one hand, online football match information is

diverse and colorful, but on the other hand users cannot find the football match information they need. Therefore, the collection, release, and related information processing of online information based on WWW have increasingly become the focus. As web information collection is playing an important role, in addition to the deepening of application and the development of technology, it is more and more used in many kinds of services and research, such as site structure analysis, page validity analysis, web graph evolution, content security detection, user interest mining, and personalized information acquisition. In short, web information collection refers to the process of automatically obtaining page information from the web through the link relationship between web pages and continuously expanding to the required web pages with the link [8].

The goal of traditional web information collection is to collect as many information pages as possible, even the resources on the whole web. This process does not care much about the order of collection and the related topics of the collected pages. One of the great advantages of this method is that it can focus on the speed and quantity of acquisition, and it is relatively simple to implement. For example, an adaptive tracking algorithm for competition moving objects was designed by Ma and Yu [9]. Moving object tracking is one of the core technologies in the field of computer vision. The implementation of software and hardware is of great significance to promote video and image processing. Taking football matches as an example, aiming at many real-time and high-precision tracking tasks of moving targets, the algorithm can only extract real-time game information. Still, it cannot collect relevant information from web pages, which is one-sided in practical application. A lightweight model retrieval algorithm for Web3D based on the SVM learning framework was proposed by Zhou and Jia [10]. This algorithm was based on sketch model retrieval. The algorithm of lightweight processing of the 3D model and selecting the best viewpoint of the 3D model based on a support vector machine was proposed. Some deep learning based methods are proposed in recent years [11, 12]; they use an end-to end way to solve the object tracking and crowd counting problems. Reference [11] presents a high-resolution network for visual recognition problems. The superior results on a wide range of visual recognition problems suggest that the proposed model in [11] is a stronger backbone for visual recognition. The poor convergence of support vector machines leads to the low accuracy of information collection, which cannot meet information collection needs in actual football matches. At present, foreign countries generally use the Google collection system to collect information on football matches. This traditional collection method also has many defects. The traditional information collection based on the whole web needs to collect many pages, which needs to consume a lot of system resources and network resources, and the consumption of these resources does not lead to a higher utilization rate of collected pages.

To effectively improve their utilization efficiency, we need to find a new way to develop a scene information collection algorithm for general football games based on web documents, break through the original traditional mode, and design a more effective scene information collection algorithm of the football game. A scene information collection algorithm of a general football match based on web documents is proposed to obtain the scene information in a general football match more comprehensively. Through the construction of a complex network and the introduction of local contribution and overlap coefficient of community discovery feature selection algorithm, web document features are selected to realize the collection of scene information in football matches. The superior results on a wide range of visual recognition problems suggest that our proposed model is a stronger backbone for visual recognition.

The rest of this paper is organized as follows. The framework and technical details of our proposed system are described in Section 2. In Section 3, we present extensive experimental results to demonstrate the effectiveness of the proposed model. Finally, we conclude our work in Section 4.

2. An Information Collection Algorithm of General Football Matches Based on Web Documents

2.1. Web Crawler Construction Based on T-Graph

2.1.1. Building T-Graph Web Crawler Model. The sample node of the specific topic of scene information in football matches is collected and linked by level. The layer where the target page is located is the zero layer, the layer linking the target page is the first layer, and so on. Repeat this process until a considerable number of nodes are established. The highest level node can directly link to the lowest level target page. There is no link between nodes at the same level. At least one link of any level node points to its lower level node. There are some special nodes in T-graph, so it cannot find the target page by calculating the similarity. Following its pointing path cannot find the target page. Such a node is called a dead node, so it needs to avoid a dead node when building a T-graph. The performance of the T-graph is tested with a known document. If the expected standard is not met, the model needs to be built repeatedly. A schematic diagram of T-graph structure is shown in Figure 1.

2.1.2. Crawling Stage of Web Crawler. Figure 2 shows the process of the crawler algorithm based on the T-graph. The event sequence of crawler crawling web page based on T-graph is as follows:

(1) The crawler selects the link with high priority from the crawling queue and sends the request to download the corresponding web page to the web network.

(2) The crawler obtains the corresponding football match information web page from the web.

(3) The crawler stores the crawled web pages in the response queue.

(4) Extract the links from the response queue and calculate the similarity with the nodes in T-graph.

RETRACTED

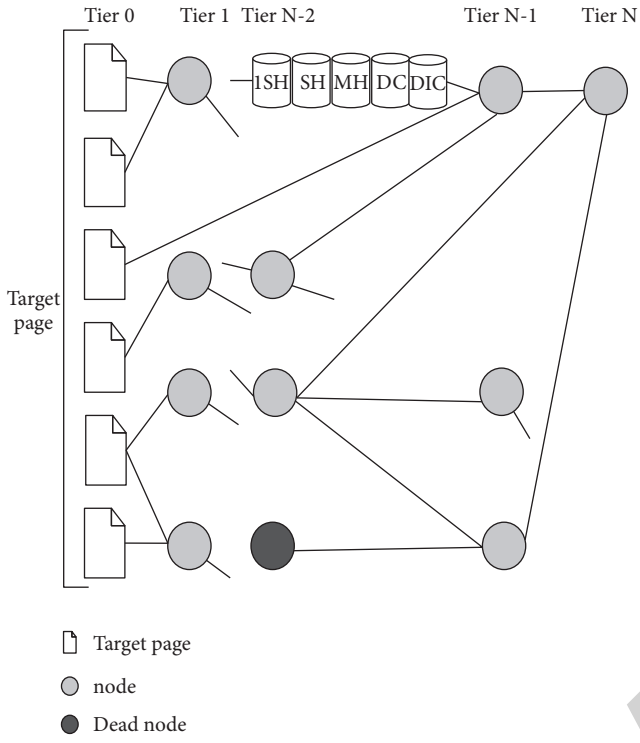

Figure 1: Schematic diagram of T-graph structure.

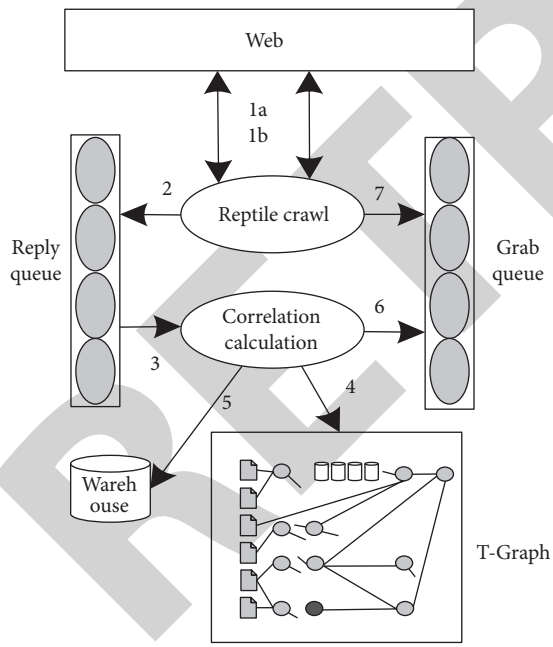

Figure 2: Crawler algorithm process based on T-graph.

(5) If the link in the web page matches the node of T-graph, the web page will be downloaded to the warehouse for storage.

(6) Extract the links in the web page and put them in the crawling queue according to the priority order.

(7) The crawler selects the link with high priority from the crawling queue for priority crawling.

The response queue stores the web page crawled by the crawler and the HTTP response. If the captured web page cannot be downloaded due to network interruption or old link, the system still maintains the details of the current HTTP response and performs similarity calculation [9, 10]. If there is no node in T-graph that matches the link in the web page, the link in the web page is still put into the crawling queue, but the link is given a lower priority. To a certain extent, this method avoids discarding the precursor nodes which are not related to the football match information topic but connected with the destination page and improves the recall.

2.1.3. Collection of Subject Edge Documents of Scene Information in Football Match. In this paper, ICTCLAS3.0 word segmentation system is used to divide the scene information in football match document into keywords. Because each keyword has one or more concepts, each keyword corresponds to one or more pieces of scene information in football match and corresponds to one or more points in two-dimensional coordinates [13]. Figure 3 shows the schematic diagram of topic edge extraction.

The circle points in the graph correspond to the keywords of the anchor document, and the triangle points correspond to the keywords of other documents. This phenomenon is called galaxy [14]. The keywords corresponding to the points in the galaxy are called topic edge documents of candidate links.

2.1.4. Information Similarity Calculation Based on Word Meaning Analysis. Considering the weight of documents in different positions, feature extraction algorithm based on semantic analysis is used to extract feature. The similarity of 1SH, SH, MH, and DC is calculated and recorded as sim_{1SH}, sim_{SH}, sim_{MH}, and sim_{DC}, respectively, and different location weights are given according to their positions in the web page. The similarity calculation formula of candidate link (CL) is as follows:

$$\text{Link sim}(CL) = P_1 * \text{sim}_{1SH} + P_2 * \text{sim}_{SH} + P_3 * \text{sim}_{MH} + P_4 * \text{sim}_{DC}, \quad (1)$$

where P_1, P_2, P_3, and P_4 can be used to plan the relevant weights of documents in different locations, and $\sum_{i=1}^{4} P_i = 2$. By increasing a certain weight, the importance of documents in corresponding locations can be increased [15], thus affecting the number of pages to be crawled. As the four attributes of 1SH, SH, MH, and DC can well distinguish topics, this paper sets $P_1 = P_2 = P_3 = P_4 = 1/4$, which means giving the same weight to the main title, section title, subtitle, and data component. The four attributes of T-graph node are all composed of documents. The document is segmented, and the feature items are extracted by TD-IDF algorithm and mapped to VSM (vector space model). The document vector P is formed. The crawled web page is decomposed structurally, and the four attributes of candidate links are extracted. After the same steps, the document vector T is formed. The similarity calculation formula of P and T is as follows:

RETRACTED

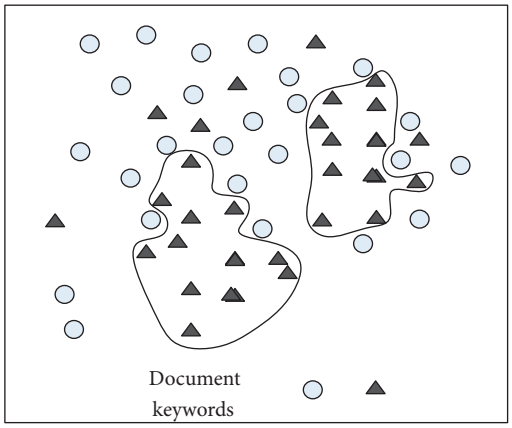

Figure 3: Schematic diagram of topic edge extraction.

$$\text{sim}(P,T) = \frac{P \cdot T}{|P||T|}. \quad (2)$$

The above formula is a mechanical matching of document keywords, which has a certain semantic deviation and affects the accuracy of similarity. On this basis, using the relevant knowledge of Wikipedia, the concept of semantic calculation and sememe of keywords is introduced to calculate the semantic similarity of candidate links from the semantic level of keywords. Concepts can be decomposed into finite sememes, and the operation of words can be transformed into the operation of sememes [16]. Suppose that document P has t_i characteristic items, expressed as n-dimension vector. By calculating the weight W of the feature items, the special effect vector representation of the document is transformed into the vector representation of the sememe set, and the weight of each feature item is given to its own sememe set, which is represented by S. After adding the weight of the same sememe set, the similarity of the sememe is calculated by calculating the similarity of the feature items. The calculation formula is as follows:

$$\text{sim}\left(P^S, T^S\right) = \cos\theta = \frac{\sum_{i=1}^{n}\left(W_i^S * K_i^S\right)}{\sqrt{\left(\sum_{i=1}^{n} W_i^{S^2}\right)\left(\sum_{i=1}^{n} K_i^{S^2}\right)}}. \quad (3)$$

Among them, W represents a multidimensional vector value, and K represents a weight value.

2.2. Document Classification Method Based on Community Discovery Algorithm. In the classification of Chinese scene information documents in football match, words are often regarded as the smallest language unit, and the amount of Chinese scene information in football match entries is very large, which makes the dimension of feature space of various classification algorithms very high. Therefore, according to the definition of complex network, each element in the system is treated as a node, and the relationship between each element is expressed as an edge, that is, a link, forming a complex relationship network. This idea of using small world features of complex network to extract key feature items provides a new idea for document feature selection. Through the discussion of complex network community structure, on the one hand, we can better understand and explain the social network presented. On the other hand, we can apply the complex network community structure theory to the specific collection of scene information in football match, which is helpful to better design the actual network function [13]. On the basis of this idea, this paper proposes a community-based document feature selection algorithm. In the process of discovering communities of the same category, the focus of document extraction is to extract the information of football match scene and train those who have strong ability to distinguish categories in the document set.

2.2.1. Community Discovery Algorithm and Complex Network Construction. Due to the uncertainty of the community discovery algorithm and in the face of a large number of nodes, it is not necessary to carry out a very strict division. This paper uses the betweenness based community discovery algorithm, namely, GN algorithm, to segment the community by removing the edge with the highest betweenness. The algorithm is as follows:

(1) Calculate all the edge betweenness in the network
(2) Remove the edge with the highest betweenness
(3) Recalculate all edge betweenness of the intermediate state
(4) Repeat (2) until all edges are removed

GN algorithm needs to analyze the whole network in every calculation, and because there is no quantitative definition of community, it is difficult to divide the community. Therefore, in order to improve the efficiency of community discovery, community division is defined as follows:

$$Q = \sum i\left(e_{ij} - a_i^2\right). \quad (4)$$

The modularization degree of complex network is measured by (4), where e_{ij} is the proportion of edges connecting communities i and j in the total number of edges, and Q is the difference between the proportion of edges falling in such communities and the expected value when the same number of edges are randomly connected. It is used to measure the modularity of complex network. The main idea of the fast algorithm based on modularization is as follows: assume that n nodes in the initial state form n communities. In the above formula, a_i is calculated as follows:

$$a_i = \sum j \cdot e_{ij}. \quad (5)$$

Using the greedy algorithm, the nodes belonging to the same kind of community are connected by continuously merging with the community whose Q value grows the fastest or decreases the slowest [17]. Radicchi et al. improved GN algorithm in 2003 and proposed a method to quantify the definition of community. Let G denote the whole network and A_{ij} be the adjacency matrix; the calculation formula of node's i degree is as follows:

RETRACTED

$$K_i = \sum j \cdot A_{i,j}. \tag{6}$$

Considering subgraph V_G, for any $i \in V$, the total degree of node i can be divided into two parts. The formula is as follows:

$$Ki(v) = K_i^{\text{in}}(v) \cdot K_i^{\text{out}}(v), \tag{7}$$

where the internal variable of V connected with node i is represented by $K_i^{\text{in}}(v)$ and the external variable of V connected with node i is represented by $K_i^{\text{out}}(v)$. The calculation formula is as follows:

$$\begin{aligned} K_i^{\text{in}}(v) &= \sum j \in vA_{ij}, \\ K_i^{\text{out}}(v) &= \sum j \notin vA_{ij}. \end{aligned} \tag{8}$$

Therefore, the definitions of strong community and weak community are given.

Definition 1. Subgraph V satisfies the definition of strong community. If and only if $K_i^{\text{in}}(v) = \sum_{j \in v} A_{i,j}, \quad j \in V$, this indicates that the number of edges connected by nodes inside the community is greater than that of edges connected by nodes outside the community.

Definition 2. Subgraph V satisfies the definition of weak community. If and only if $\sum i \in vK_i^{\text{in}} > \sum i \in vK_i^{\text{out}}$, this means that the sum of the number of connecting edges between all nodes in the community and nodes inside the community is greater than the sum of the number of connecting edges between all nodes and nodes outside the community.

(1) Choose a way to define a quantitative community.

(2) Calculate the betweenness of all edges and remove the edge with the largest betweenness.

(3) After removing the edge, if the network is not divided into two parts, repeat (2).

(4) If the removed edge can divide the network, judge whether there are at least two subnets meeting the definition of quantitative community selected in step (1). If so, mark the corresponding part on the graph.

(5) Return to step (2); all subnetworks continue to execute until there is no edge in the network. By constantly removing the edge to construct the community and quantifying the community, the community division is more reasonable.

2.2.2. Feature Selection Algorithm of Scene Information in Football Match. For GN algorithm, when the number of nodes exceeds thousands, the computational complexity will become very high, and there is no quantitative definition of community. If we directly use fast algorithm to aggregate items of the same category, the accuracy is difficult to guarantee. Therefore, based on the consideration of time complexity and accuracy, the idea of local contribution and overlap coefficient is introduced, which is called feature selection algorithm based on community discovery. Firstly, according to the definition of community, the definition of Q value is different from that in this paper. The formula is as follows:

$$Q = k_i^{\text{in}}(V) k_i^{\text{out}}(V). \tag{9}$$

It represents the difference between the number of edges connected between a node i and the internal node of community V and the number of edges connected by the external node of the community. The main idea of the algorithm can be described as follows: through each category community of the initial preclassification, the feature nodes in the complex network that meet the definition of each category community are selected [18]. The specific algorithm is as follows:

(1) The complex network graph is constructed based on the training text set.

(2) Initialization community: The elements in the set $S = \{C1, C2, \ldots, Cn\}$ are predefined by experts, and each community represents a category. The community is composed of a small number of feature nodes (generally 10-20, 20 in this section) of each category with a strong ability to distinguish categories. Each feature node, except the feature node in the predefined community, constitutes a community setting in the network. The set expression is as follows:

$$T = \{C_{n+1}, C_{n+2}, \ldots C_{n+m}\}. \tag{10}$$

(3) For each community C_k in the set S, the Q values of C_k and C_j are calculated, respectively, and the Q values are arranged in descending order. The first 10 C_j and C_k are merged into a community. If the first 10 C_j have Q values less than 0, only C_j and C_k greater than 0 are merged, the merged nodes are removed from the set T, and the newly added nodes in each predefined community are recorded.

(5) After several steps, the new nodes added to each predefined community are checked according to the strong definition of the community, and the nodes that do not meet the conditions are deleted. The nodes deleted for the first time are not permanently deleted but are used by the next community. If the same node is deleted for the second time, the node is permanently deleted.

(6) Return to step (3); the number of nodes in each predefined community should meet the number of features selected, or there are nodes whose Q is less than or equal to 0 that can join a predefined community.

In the experiment, it is found that there is an overlapping phenomenon in the partition of edge and intermediate points. Accordingly, the algorithm is improved as follows:

Improvement (1): According to the idea of local contribution degree, the largest node of degree (central node) is taken as the initial community, and then the neighbor points (the former neighbor points with stronger differentiation) which have the greatest contribution to the community are

RETRACTED

added in turn. When the contribution degree reaches extreme value, a community can be formed. If there are multiple boundary nodes with large contributions, they are added to multiple communities sharing it. After the community is extracted, the nodes and edges of the community are not deleted from the network to facilitate the mining of edge mediators [19].

Improvement (2): By limiting the overlap coefficient, if the overlap coefficient of C_i and C_j in any two communities is greater than the threshold T, the merged community becomes a whole (T is taken as 0.7 in this paper). At this time, the local contribution calculation formula is as follows:

$$q = \frac{l_{\text{in}}}{l_{\text{in}} + l_{\text{out}}}. \tag{11}$$

In (11), l_{in} represents the number of links within the community, l_{out} represents the number of links outside the community, and the greater the Q value is, the greater the contribution to the community is. The global contribution degree Q represents the current maximum contribution degree in the mining process, which is initialized to 0 and used to judge whether the current community has reached the best state [20]. The overlap coefficient S is calculated as follows:

$$S = \frac{\left|C_i \cap C_j\right|}{\left|C_i \cup C_j\right|}. \tag{12}$$

In the above formula, the numerator represents the number of common nodes of communities C_i and C_j, the denominator represents the number of all nodes of C_i and C_j, and the set of adjacent points is marked as U. The implementation basis of the feature selection algorithm of scene information in football match based on community discovery is complex semantic network graph, and the threshold value of the algorithm is 0.7. When dividing communities, when the threshold value is greater than 0.7, the two overlapping communities are merged. The specific flow of the algorithm is shown in Figure 4.

3. Experimental Analysis

We chose the scene information in a general football match in 2015–2019 as the theme for the test, collected 50 information theme websites of football match, and added 100 unrelated websites to form the test set, which contains more than 80000 pages. The measurement index is utilized to evaluate the topic collection efficiency of this algorithm comprehensively. Experiments in this paper are carried out using one GPU (GeForce GTX 1050 Ti) and an Intel Core i7 with 16 GB RAM system. We have added the hardware enlivenments in the revised manuscript.

The accuracy of acquisition is defined as follows: the number of theme related pages in collected pages/the number of all collected pages.

The resource discovery rate is defined as follows: the number of pages related to topics in collected pages/the number of pages related to all topics.

We use the same set of scene information in football matches to collect data. To effectively get the accurate effect of each method, we suspended the page and topic correlation determination module in the experiment. In the experiment, the number and status of pages when the number of collected pages is 1000, 2000, 3000, ..., 10000, respectively, are recorded, and the collection accuracy and resource discovery rate are calculated in time. When calculating the collection accuracy and resource discovery rate, we must know how many pages are related to the topic. Although the accuracy of this method is not as accurate as that of the manual method, the automatic determination of the machine saves a lot of time. In this paper, the algorithms in [9–11] are used to test the acquisition accuracy and resource discovery rate. The results are shown in Table 1. The algorithm in [9] represents an adaptive tracking algorithm for competition moving objects, which proposes a new adaptive target tracking algorithm based on feature fusion and particle filter. The algorithm hardware platform based on an image processing unit is designed. The algorithm in [10] studies the related technologies of 3D model retrieval based on sketch and puts forward the lightweight processing algorithm of 3D model and the optimal viewpoint selection algorithm of 3D model based on support vector machine. Reference [11] presents a high-resolution network for visual recognition problems. The superior results on a wide range of visual recognition problems suggest that the proposed model in [11] is a stronger backbone for visual recognition.

Analysis of Table 1 shows that, with the increase of data volume, the three algorithms' resource acquisition accuracy and resource discovery rate also decrease. The decline of the resource acquisition accuracy and resource discovery rate of the text algorithm is lower than that of the algorithm in [9] and the algorithm in [10]. When the data volume is 10000, the resource acquisition accuracy of the algorithm in this paper is 7.7% and 10.41% higher than that of the algorithm in [9] and the algorithm in [10], respectively. In comparison, the resource discovery rate of the algorithm in this paper is 15.52% and 19.13% higher than that of the algorithm in [9] and the algorithm in [10], respectively, and average resource acquisition accuracy and resource discovery rate of the algorithm in this paper are 97.46% and 97.68%, respectively. The algorithm's average resource acquisition accuracy and resource discovery rate in [9] are 94.11% and 83.93%, respectively. The algorithm's average resource acquisition accuracy and resource discovery rate in [10] are 91.44% and 83.98%, respectively. The comprehensive comparison shows that the algorithm in this paper has a high ability to collect topics.

The algorithm cost and acquisition accuracy of the three algorithms are compared, and the results are shown in Table 2.

The qualitative comparison in Table 2 shows the following: As for cost,the cost of the algorithm in this paper is the smallest, which is equivalent to not doing any similarity calculation and comparison. However, the algorithms in [9] and [10] only compare the extended metadata in each link because the amount of information in the extended metadata is minimal, and the cost of time and space is minimal.

RETRACTED

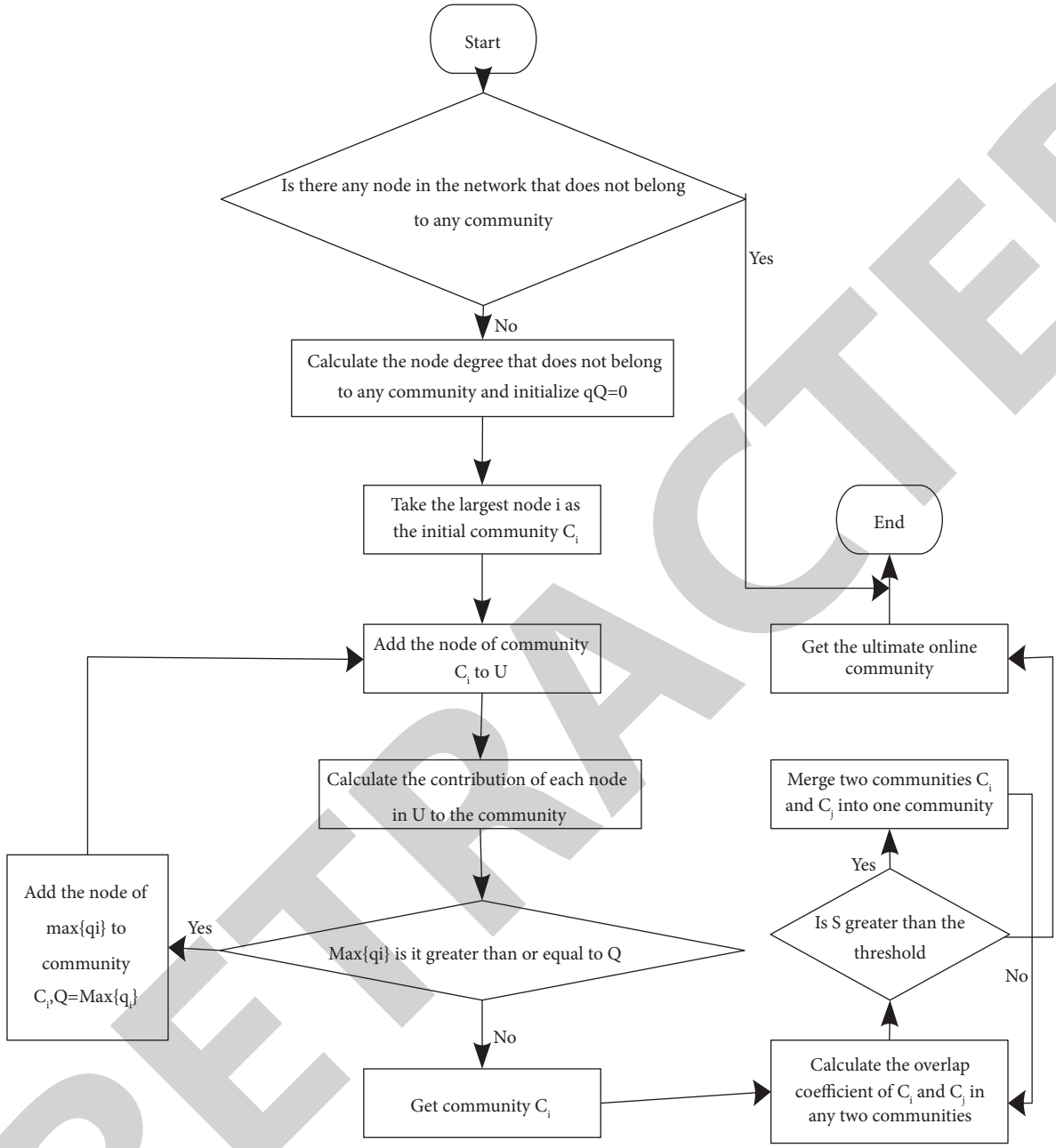

Figure 4: Flow chart of feature selection algorithm for scene information in football match based on community discovery.

Still, it is more complex than that of the algorithm in this paper. For the algorithm of [9] and the algorithm of [10], when the characteristics of the significant pages are found, the critical pages are collected first, and the collection accuracy increases to a certain degree. When the quality of the pages is not too high, the collection accuracy decreases. Therefore, the algorithm in this paper has a small calculation cost and low impact on collection accuracy.

Mean average precision (mAP) is used to measure the performance of the algorithm. The average precision (AP) of a certain category is the sum of the precision of different recall test points divided by the number of recall test points. For the entire dataset, mAP is the sum of APs of all categories divided by the number of categories. The crawler can identify the relatively high priority link of the football match site information in the test and can obtain the corresponding site information web page of football match from the web. If the recognition error is less than 0.3 s, it is considered that the recognition is correct. In the process of AP calculation, 8000 information web pages of football matches are used as test points to calculate recall rate and accuracy, further calculate the balanced average accuracy mAP, and calculate the frame rate of the three algorithms. The calculation results are shown in Figure 5.

It can be seen from the analysis of Figure 5(a) that, with the increase of the number of web pages, the mAP of the

RETRACTED

Table 1: Test results of resource collection accuracy and resource discovery rate of three algorithms.

Data volume	The algorithm in this paper		Reference [9] algorithm		Reference [10] algorithm		Reference [11] algorithm	
	Acquisition accuracy (%)	Resource discovery rate (%)	Acquisition accuracy (%)	Acquisition accuracy (%)	Resource discovery rate (%)	Resource discovery rate (%)	Acquisition accuracy (%)	Resource discovery rate (%)
1000	**99.66**	**99.87**	99.03	95.46	91.43	89.57	97.56	95.46
2000	**99.32**	**99.58**	98.47	95.37	90.41	88.24	96.39	98.43
3000	**98.79**	**99.01**	97.55	94.55	88.74	86.41	96.57	93.78
4000	**98.01**	**98.76**	96.47	93.33	86.53	85.61	95.38	93.56
5000	**97.22**	**98.16**	95.77	92.71	84.31	85.03	96.78	95.36
6000	**97.16**	**97.49**	93.24	91.17	83.33	83.22	95.19	96.35
7000	**96.54**	**97.09**	92.22	90.22	80.54	81.45	94.62	93.64
8000	**96.11**	**96.57**	90.47	88.54	79.99	80.73	92.56	89.89
9000	**95.99**	**95.46**	89.76	87.62	78.91	79.82	91.66	88.96
10000	**95.84**	**94.77**	88.14	85.43	75.64	79.25	89.48	87.69

three algorithms presents a downward trend. The mAP of the algorithm in this paper decreases slightly, and the mAP is always maintained at about 98%. In contrast, the mAP of the algorithms in [9] and [10] decreases greatly. When the number of web pages is 8000, the mAP of the two algorithms is 63% and 69%, respectively, which is quite different from the algorithm in this paper. It can be seen from the analysis of Figure 5(b) that, in the process of web page calculation, the number of frame rates fluctuates to a certain extent. The number of frame rates calculated by the algorithm in this paper is always high and remains between 35 and 40. In contrast, the algorithms in [9] and [10] fluctuate greatly, and the number of frame rates calculated spans a large range. Comprehensive analysis of Figure 5 shows that the algorithm in this paper has high average accuracy and fast calculation speed.

The web crawling ability of the three algorithms is tested. The web crawling condition is set as follows: the update cycle of a web page is 10 min/time, five times every three hours, with continuous crawling for 10 hours. Each crawling only retains the effective scene information in football matches. If the captured scene information in the football match has been saved in the warehouse, the web page will be discarded. The results of the three algorithms are shown in Figure 6.

As shown in Figure 6, the algorithm in this paper crawls about 58000 web pages in the first crawling cycle because, in this stage, the web page is crawled for the first time, so most web pages are retained. The second peak of crawling occurs after 3 hours (because the interval of crawling is 3 hours), and about 15000 web pages are picked up. After that, the number of web pages crawled every day tends to be flat. After one crawling cycle, the access pages can be basically recognized with focus on capturing the data and related links. At the same time, the peak of the algorithms in [9] and [10] is not obvious, and the number of web pages is small. Therefore, the algorithm in this paper has a strong ability to capture web pages.

In the test, the number of lost packets to the number of sent data groups is the packet loss rate. The packet loss rate is closely related to the packet length and the packet transmission frequency. The calculation results of the three algorithms are shown in Figure 7.

Analysis of Figure 7 shows that, with the increase of the number of web pages, the packet loss rate of the three algorithms also increases gradually. When the number of web pages is 8000, the packet loss rate of the algorithm in this paper is only 3%, which is 3% and 5% lower than the algorithms in [9] and [10], respectively, with the difference being relatively large.

The 8000-page scene information in football matches is divided into 8 communities. The three algorithms are used to calculate the edge betweenness, and their community quantization ability is tested. The results are shown in Figure 8.

It can be seen from the analysis of Figure 8 that although the calculation of edge betweenness fluctuates with the increase of the number of communities, the value of edge betweenness calculated by the algorithm in this paper is the largest. The edge betweenness calculated by the algorithm in [9] is similar to that calculated by the algorithm in this paper before the number of communities is 2. However, with the increase of the number of communities, the edge betweenness calculated by the algorithm in [10] decreases greatly. The computing edge betweenness is always low, so the algorithm in this paper has a strong community quantification ability.

The data collection ability of the three calculation methods is tested. Data collection was performed 5 times per hour, and the amount of scene information in football matches collected by the three algorithms after 12 consecutive periods was counted. The results are shown in Table 3.

According to the analysis of Table 3, the amount of scene information in football matches collected by the algorithm in this paper is 48682, which is higher than the algorithms in [9], [10], and [11], respectively. When the collection period is 3, 5, 8, and 10, the moving target adaptive tracking algorithm proposed in [9] has a better collection number, because the algorithm has good adaptive ability, and it just has a certain period of time. However, in terms of the number of collections in the entire 12 cycles, the method in this article has significant advantages. It can be seen that the algorithm in this paper has a strong information collection ability.

RETRACTED

Table 2: Changes of algorithm cost and acquisition accuracy of the three algorithms.

Algorithm	Algorithm cost	Change of acquisition accuracy
The algorithm in this paper	Less	The accuracy is high in the early stage, and then it begins to decline slowly
Reference [9] algorithm	Moderate	As time goes on, the rate of change is not big, but it rises suddenly occasionally
Reference [10] algorithm	Large	With the increase of time, the accuracy curve first increases, then decreases, and fluctuates greatly

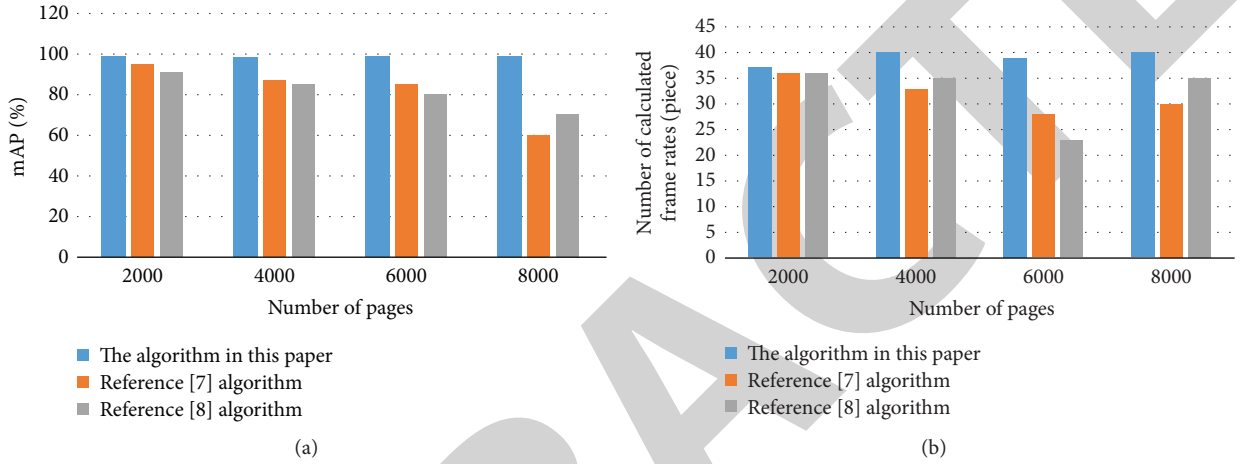

Figure 5: Test results of the three algorithms: (a) mAP test results; (b) calculation speed.

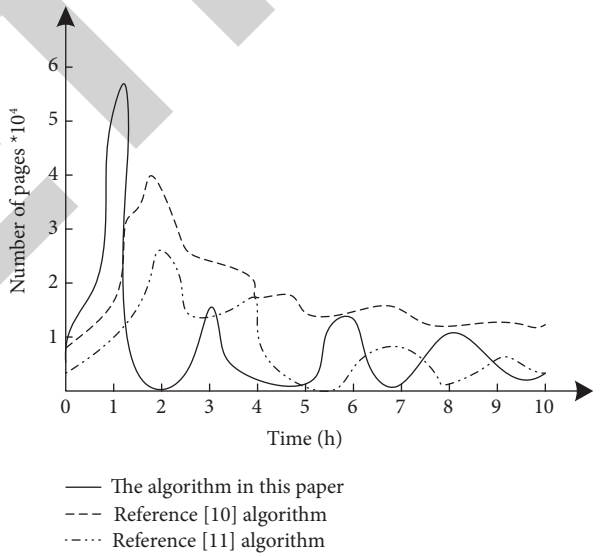

Figure 6: The results of the three algorithms for web crawling.

RETRACTED

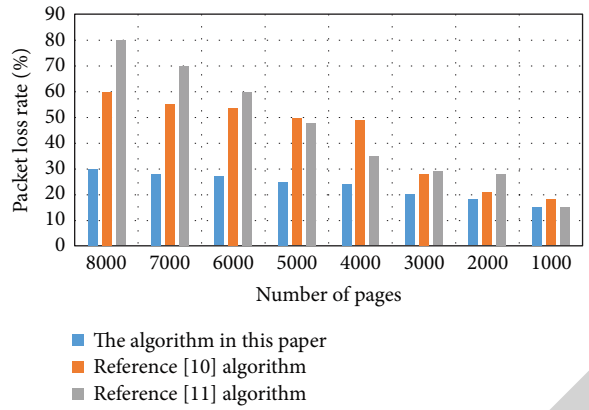

Figure 7: Packet loss rate test results of the three algorithms.

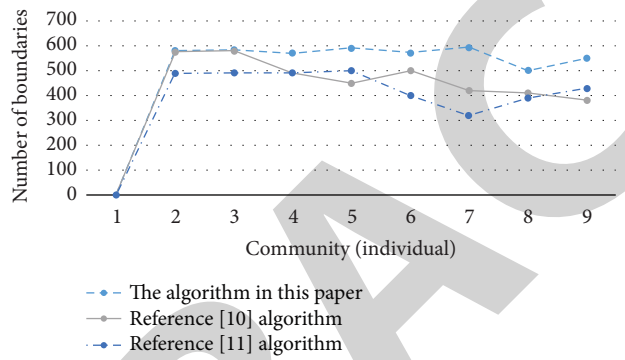

Figure 8: Test results of community quantification ability of the three algorithms.

Table 3: Statistics of the amount of scene information in football match collected by the three algorithms in 12 periods.

Collection period	The algorithm in this paper	Reference [9] algorithm	Reference [10] algorithm	Reference [11] algorithm
1	**3889**	2363	3412	3615
2	**4323**	2543	2199	3120
3	4988	**5017**	2076	4086
4	**5012**	3324	3407	4408
5	3124	**3501**	5511	4502
6	**4096**	3896	3076	3023
7	**4327**	4227	4134	4021
8	3914	**5011**	3983	3863
9	**5339**	2987	3042	4056
10	2176	**3512**	2206	2023
11	**3073**	2983	3863	3103
12	**4421**	4307	3147	3657
Total	**48682**	43671	40056	43477

4. Conclusion

With the continuous improvement of network service types and quality requirements, this new idea of data collection has emerged. For this reason, we propose an information collection algorithm for general football matches based on web documents. The introduction of web documents in target information prediction and collection helps to realize personalized intelligent service. Personalized active information collection service has become a hot spot that people pay more and more attention to, and it is a development trend of collection service in the future. With the continuous improvement of its function, accuracy, and intelligence, the personalized prediction collection mode based on users will play a more important role and better meet users' needs. The topic crawler strategy based on T-graph, by analyzing the topic edge text of candidate links, predicts the correlation between links and topics, comprehensively considers the page content and link analysis, and improves the quality of theme crawling. The experimental results show that the proposed method is better than the baselines in practical applications. Generally speaking, the algorithm is successful and meets the expected requirements of scene information collection in general football matches. However, there is still

RETRACTED

a lack of reasonable theoretical analysis of our method. In the future work, we will discuss the models and methods in the analysis in a theoretical way [21–25].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful for the grants provided by the National Social Science.

References

[1] L. S. Hu, "Design of web collection algorithm based on lda model," *Journal of Daqing Normal University*, vol. 038, no. 6, pp. 55–58, 2018.

[2] M. Li, N. Cheng, J. Gao, Y. Wang, L. Zhao, and X. Shen, "Energy-efficient UAV-assisted mobile edge computing: resource allocation and trajectory optimization," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 3424–3438, 2020.

[3] P. Cong, J. Zhou, L. Li, K. Cao, T. Wei, and K. Li, "A survey of hierarchical energy optimization for mobile edge computing," *ACM Computing Surveys*, vol. 53, no. 2, pp. 1–44, 2020.

[4] C. Lin, G. Han, X. Qi, M. Guizani, and L. Shu, "A distributed mobile fog computing scheme for mobile delay-sensitive applications in SDN-enabled vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5481–5493, 2020.

[5] S. Agarwala and S. Fryer, "An algorithm to construct the le diagram associated to a grassmann necklace," *Glasgow Mathematical Journal*, vol. 62, no. 1, pp. 85–91, 2019.

[6] G. Cheng, "Tret: a text retrieval efficiency testing tool for different document types/formats and calculating evaluation measures for xml retrieval," *Advances in Computational Sciences and Technology*, vol. 11, no. 3, pp. 233–246, 2018.

[7] G. Han, S. Shen, H. Wang, J. Jiang, and M. Guizani, "Prediction-based delay optimization data collection algorithm for underwater acoustic sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 99, pp. 6926–6936, 2019.

[8] T. Gayathri and D. L. Bhaskari, "A novel scalable signature based subspace clustering approach for big data," *International Journal of Information Technology and Web Engineering*, vol. 14, no. 2, pp. 41–51, 2019.

[9] Y. J. Ma and L. Yu, "GPU implementation of an adaptive tracking algorithm for competition moving target," *Natural Science Journal of Xiangtan University*, vol. 40, no. 4, pp. 71–74, 2018.

[10] W. Zhou and J. Y. Jia, "A lightweight model retrieval algorithm for Web3D based on SVM learning framework," *Acta Electronica Sinica*, vol. 47, no. 01, pp. 92–99, 2019.

[11] J. Wang, K. Sun, T. Cheng et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, 2020.

[12] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: a large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2141–2149, 2020.

[13] H. Zhang, Z. Li, W. Shu, and J. Chou, "Ant colony optimization algorithm based on mobile sink data collection in industrial wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 152, 2019.

[14] W. Zhao, J. Luo, S. Li, J. Qi, H. Meng, and Y. Li, "Design of dynamic calf weighing system based on moving-iir filter algorithm," *Journal of Electrical Engineering & Technology*, vol. 16, no. 2, pp. 1059–1069, 2020.

[15] C. Botezatu, B. Mastalier, and T. Patrascu, "Hepatic hydatid cyst – diagnose and treatment algorithm," *Journal of medicine and life*, vol. 11, no. 4, p. 394, 2018.

[16] N. Mittal, U. Singh, R. Salgotra, and B. S. Sohi, "An energy efficient stable clustering approach using fuzzy extended grey wolf optimization algorithm for wsns," *Wireless Networks*, vol. 25, no. 8, pp. 5151–5172, 2019.

[17] Y. Ran, Y. Shi, C. Tang, and Z. Zhang, "A primal-dual algorithm for the minimum partial set multi-cover problem," *Journal of Combinatorial Optimization*, vol. 39, no. 3, pp. 725–746, 2020.

[18] L. Patrick, S. Moses, K. Joanne, and F. E. Robles, "Dual-wavelength oblique back-illumination microscopy for the non-invasive imaging and quantification of blood in collection and storage bags," *Biomedical Optics Express*, vol. 9, no. 6, p. 2743, 2018.

[19] M. Abdallah, M. Adghim, M. Maraqa, and E. Aldahab, "Simulation and optimization of dynamic waste collection routes," *Waste Management & Research*, vol. 37, no. 8, pp. 793–802, 2019.

[20] Y. Miao, Z. Sun, N. Wang, Y. Cao, and H. Cruickshank, "Time efficient data collection with mobile sink and vmimo technique in wireless sensor networks," *IEEE Systems Journal*, vol. 12, no. 1, pp. 639–647, 2018.

[21] K. Liu, L. Xu, and J. Zhao, "Co-extracting opinion targets and opinion words from online reviews based on the word alignment model," *Knowledge & Data Engineering IEEE Transactions on*, vol. 27, no. 3, pp. 636–650, 2018.

[22] J. Wang, S. F. Tan, D. D. He, J. H. Chen, and J. Z. Yan, "Entity disambiguation algorithm for domain document based on context feature," *Beijing Biomedical Engineering*, vol. 37, no. 4, pp. 398–402+409, 2018.

[23] Z. Qin, A. Li, C. Dong, H. Dai, and Z. Xu, "Completion time minimization for multi-uav information collection via trajectory planning," *Sensors*, vol. 19, no. 18, p. 4032, 2019.

[24] N.-T. Nguyen, B.-H. Liu, V.-T. Pham, and T.-Y. Liou, "An efficient minimum-latency collision-free scheduling algorithm for data aggregation in wireless sensor networks," *IEEE systems journal*, vol. 12, no. 3, pp. 2214–2225, 2018.

[25] C. B. Yin, S. Q. Sun, and S. L. Yi, "Operation and maintenance optimization simulation of power consumption information collection based on big data analysis," *Computer Simulation*, vol. 35, no. 011, pp. 436–439, 2018.

RETRACTED