Hindawi
Complexity
Volume 2024, Article ID 9758761, 1 page
https://doi.org/10.1155/2024/9758761

Retraction

Retracted: Internet Tourism Resource Retrieval Using PageRank Search Ranking Algorithm

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets 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 own 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] H. Li, "Internet Tourism Resource Retrieval Using PageRank Search Ranking Algorithm," *Complexity*, vol. 2021, Article ID 5114802, 11 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 5114802, 11 pages
https://doi.org/10.1155/2021/5114802

Research Article

Internet Tourism Resource Retrieval Using PageRank Search Ranking Algorithm

Hui Li

School of Tourism, Guangdong Polytechnic of Science and Technology, ZhuHai 519090, China

Correspondence should be addressed to Hui Li; luck0756@jluzh.edu.cn

Received 24 April 2021; Revised 10 May 2021; Accepted 19 May 2021; Published 27 May 2021

Academic Editor: Zhihan Lv

At present, there is a wide variety of tourism resources on the Internet. Tourism management departments must monitor these resources. At the same time, tourists must also retrieve personalized information that they are interested in. This requires a lot of time and energy. This essay studies and implements the tourism network resource monitoring system. The main work completed in the thesis proposes and constructs a topic collection algorithm and establishes a starting point, topic keywords, and a prediction mechanism. The algorithm includes three stages: the first climbing stage, the learning stage, and the continuous climbing stage. Open category directory search is used for similarity judgment and result evaluation. The experimental results show that with the continuous execution of the crawling process, the collection speed of related pages is getting faster and faster. We propose an algorithm for the extraction of wood based on the density of Internet tourism resources. The algorithm calculates the ratio of Internet tourism resource labels by row and uses a threshold extraction algorithm to distinguish area from private non-Internet tourism resource area. Experimental results show that the algorithm can successfully extract the main content of the article from a wide variety of web pages. This thesis takes the monitoring of tourism network resources as the research object and establishes a tourism network resource monitoring system, which can provide users with customizable, all-round, and real-time tourism network resource collection, extraction, and retrieval services so as to monitor tourism resources. The research results of this article can promote the construction of tourism informatization and can help users grasp the latest tourism information, thereby bringing great convenience to tourism. The system only downloads travel-related information through the use of topic collection technology, reducing the interference of irrelevant redundant web pages.

1. Introduction

Internet applications have penetrated into my country's cultural, economic, political, and social life and other fields, and China's tourism information industry has also developed rapidly [1]. The network has gradually evolved from a convenient communication tool and efficient new media to a huge virtual society [2]. The rapid development of the tourism economy and information technology has caused tremendous changes in the information-intensive industry of tourism [3]. As an important channel to provide tourism information resources, tourism websites have gradually become the main source of reference for most potential tourists to obtain information before they travel, and they have played an increasingly significant role in the travel decision-making of tourists [4]. In all subjects, travel websites have a huge amount of travel information about scenic spots, user comments, scenic spots introduction, and other related information. It takes a lot of time for tourists to extract tourist information that they are interested in from these websites [5]. Due to the business cooperation relationship, travel websites will only provide travel information that has a cooperative relationship with them, and it is difficult to provide tourists with all-round, massive, high-quality, and low-cost services [6].

In the past ten years, the technology of extracting Internet travel resources from web pages has been extensively studied, and many methods have emerged. Patel et al. [7] proposed a

mechanism using artificial intelligence to identify noisy data such as border advertisements and redundant irrelevant links. However, this technology is not suitable for practical use because it requires a huge artificially defined training set and requires knowledge of related fields to establish classification rules. Gayar et al. [8] proposed another extraction technology based on vision. Based on the algorithm, Marine [9] used the method of machine learning to sort the blocks in the web page by importance, and the sorting is mainly based on the location and size of the space attribute, the number of content attributes, pictures, and links. Gleich and Rossi [10] proposed a technique for extracting templates from custom controls contained in web pages. Chung et al. [11] proposed the structure of the website type tree, which treats similar types in the tree as meaningless. Elbarougy et al. [12] proposed an extraction algorithm to improve the accuracy of the content classification of the digital library. In order to solve the defect that the algorithm can only identify a single Internet tourism resource segment, the Internet tourism resource slope is proposed. Granka [13] proposed a link threshold filtering algorithm, which removes advertising links and navigation elements by calculating the ratio of text in hyperlinks. This technology mainly relies on the block technology of web page; the segmentation of web pages is mainly based on the location of Internet tourism resources, pictures, and scripts [14]. Then, different extraction algorithms were mixed for Internet tourism resource extraction, and the results proved that the specially selected hybrid extraction algorithm is better than the extraction algorithm alone [15]. If the starting point cannot well guide the search ranking to the relevant pages, then the number of relevant pages found by the search ranking will be very small [16]. For proposing the search and sorting system, which is developed, the system does not need to start in advance, but it can still find pages related to the topic [17]. One is proposed by and to provide a search sorting keyword describing the user's interest. Crawlers use these keywords to find candidates through search engines and start with those found. The advantage of using this technology is that users do not need relevant professional background knowledge [18]. However, if there is no relevant interest classification in the public network catalog, then the algorithm will lose its effectiveness. The topic-related crawler simply chooses a direction to visit the Internet [19]. At present, there are many sorting technologies, which can be divided into two types: link-based sorting and content-based sorting. Backlinks indicate the number of links that point to the same link. The higher the value is, the greater the importance is [20]. Forward links indicate the number of links sent from one. The page rank is the ratio of the sentences of backward links and forward links. Experiments show that web page rank is the best evaluation parameter in the ranking [21]. If the information does not exist, then the page level cannot be calculated. The concept of "hub value" is proposed in the adjacent ordering. A good hub is most suitable as a starting point because it will point to more topic-related pages [22]. Similar to the calculation process of web page rank, the pivot value also needs the link information between web pages to be calculated, then a point of view is put forward, and most of the topic-related pages are in the same parent directory. A similar view was also put forward. Pages under the same web directory are more relevant to the same topic [23] and put forward an algorithm, which can let the search sorting learn, store, and point to the path of the relevant page. For the content-based sorting algorithm, this algorithm uses the topic similarity space vector for ranking operation. The algorithm first calculates the similarity between topic keywords and web page text content [24, 25]. If the collected pages have a high degree of similarity, then this page and the pages in this page will be considered related to the topic [26–29]. Similarity includes two aspects: the similarity of content Internet tourism resources and the similarity of anchor Internet tourism resources [30, 31]. The content text similarity indicates the similarity between the content and the topic, and the anchor text similarity indicates the similarity between the web page and a certain topic [32–34].

At the same time, due to the concealment and freedom of the Internet, the Internet also contains a lot of false travel information, causing many tourists to suffer a certain degree of economic loss. Information retrieval services have penetrated into social life and brought great convenience to people's lives [35]. However, the search service dedicated to tourism is still in the exploratory stage. This article takes the tourism industry as the main object and adopts information retrieval-related knowledge to establish a tourism network resource monitoring system to provide users with customizable, omnidirectional, and real-time information delivery [36]. An improved algorithm is proposed to calculate the personal characteristic matrix, and the improved algorithm is compared with the existing algorithm. A search ranking algorithm based on scoring is used for ranking.

2. Construction of Internet Tourism Resources Retrieval Model Based on PageRank Search Ranking Algorithm

2.1. Hierarchical Distribution of Tourism Resources Retrieval. Feature matrix M is constructed according to the user's travel information retrieval history, and the category matching is performed through the user's travel information retrieval words. This article introduces the Rocchio batch learning algorithm and aims at the algorithm when there are too many retrieval records. For problems such as the low operating efficiency of the algorithm, an adaptive search strategy is used to optimize its operating efficiency. When the user enters a different search keyword, the user's search characteristics will be adaptively modified accordingly. The improved adaptive PageRank algorithm proposed in this article is shown in

$$p(x) = \sum \operatorname{sim}(x, x-i) \times y(x, x-i). \tag{1}$$

Among them, M represents the personal characteristic matrix obtained at t time, and i represents the data obtained from the 0 time to the t time and related to the retrieval category, which represents the weight of the j word in the data related to the retrieval category i obtained between the $t-1$ and t times with the following:

$$\operatorname{sim}(x, y) = \frac{\sum w(i, j)}{\sqrt{w(i) \times w(j)}}. \tag{2}$$

In order to further improve the matching efficiency, this article proposes a hybrid feature threshold extraction matching method. The hybrid feature uses user retrieval features and general retrieval features. Among them, C represents the user search feature category, and C-g represents the general search feature category. The matching algorithm for each category is as follows:

$$f(x) = \frac{n(i, j)}{\sum n(i) \times x(i, j)}. \tag{3}$$

After the user enters a search term, it is matched with the characteristics of different categories, and the three search results with the highest similarity are returned to the user. Define the searched Internet tourism resource that has not been classified by the search feature as N, and the total number of data is M. Define the data that have been archived by category and are consistent with the user search category i as N, and the total number of corresponding data is M-i; then, the results obtained from N and N retrieval will be sorted in a mixed manner.

$$m(i, j) = \frac{1}{n \times \left(\sum f(k, j) \times y(k, i)\right)}. \tag{4}$$

For each search of the user, the algorithm will feed back 3 categories with high relevance to the search term to the user and use the formula to score the relevance of the search category and the search keyword. Vector space model is a statistical model used to calculate the relevance of web pages. In this statistical model, a set of linearly independent basic vectors are used to represent web pages in the WWW. In the vector space model, in order to facilitate understanding, we use the following way: (W1, W2, ..., Wn) represents a group of web pages; (T1, T2, ..., Tn) represents the number of web pages. The feature item $Wi = (wi1, Wi2, ..., Win)$ represents the weight of the feature item in the web page; for example, the weight of the feature item T-j in the Internet tourism resource W-i is W-j. The correlation between web pages is W-i.

According to the Rel (W-i, w-j), it can be seen from Figure 1 that the VSM model uses the cosine of the angle between vectors W-i and W-j to calculate the correlation; that is, the larger the angle between vectors W-i and W-j, the less relevant of the corresponding w curve pages. Assuming that a piece of data appears more frequently in different retrieval result lists, the score of the data can be expressed as the sum of the score values of each retrieval queue.

$$s(i, j) = \frac{s(i, j \times m) + s(i \times n, j) + s(i \times j, n)}{3}, \tag{5}$$

where n represents the number of all search categories related to the search keywords, and sc ore-c represents the scores of the top three search categories c in terms of relevance. Rankci represents the ranking of the retrieval category c, and ideal_rank represents the highest possible ranking of the retrieval category c. Among them, M is the topic degree of the word adjoining in web page, T is the total number of words in eeb page j, and level (M) is the word frequency of the word in web page j. The topic degree of a word determines the importance of the word in the web page, and the topic degree can reflect the topic content of the web page. In fact, the idea of keyword thematic degree and the deterioration of word frequency are conceivable to a certain extent, and they are all developed on the basis of word frequency.

$$u(i, j) = \max(s(i \times n, j), s(i, j \times m)). \tag{6}$$

U represents the set of web pages that needs to be judged, and the vector m and the vector n represent the pivot value and core value of the page. First, the the vector m and the vector n are initialized, and the range of the core value and the pivot value is one. It indicates the core weight of the page and the pivot weight of the page. If there are links on the page of the first host, these links point to a certain page of the second host, then each link is assigned a value, and this value is used to calculate the core value of the page in the second host. In the same situation, if a web page in the first host is pointed to by a page in the second host, then each linked page is assigned a value of certainty. The core value and pivot value are used to solve mutually reinforcing problems.

2.1.1. PageRank Search Ranking Algorithm. Assuming that the length of all search result lists is N, the score of the i data in the list is $(N - i + 1)$, so the highest score of the first data in the search result list is N, and the last data have the lowest score. Assuming that a piece of data appears more frequently in different search result lists, the score of the data can be expressed as the sum of the score values of each search queue. Then, the data that appear in multiple search result lists have a higher score than the data that appear alone. First, score each retrieval result data to get the total score, and then aggregate the data appearing in the different retrieval result queues into a list and sort them in ascending order of weight. The scoring base W-j of the retrieval queue is shown in

$$v = \frac{1}{1 + t(i, j) - s(i, j) \times t(i \times a, j \times b)}. \tag{7}$$

Among them, a represents a set of keywords related to the topic and b represents a web page for comparison. Claw and carve separately indicate the number of words in the collection and the number of words in the web page. It can be seen that the similarity result is between 0 and 1. The higher the result value, the higher the similarity. This algorithm is a clustering algorithm for Internet tourism resources that has nothing to do with the content of web pages. First, we need to construct a two-way graph of hyperlinks between keywords and pages. The construction principle is as follows: all keywords are represented by circular nodes, and all hyperlinks are used. The square node indicates that if the user enters keyword A in the query interface for the query process, hyperlink B appears in the returned result page and is effectively clicked by the user; then, a two-way edge is established between keyword A and hyperlink B, represented by a solid double-headed arrow. If the hyperlink in the return result page is clicked by the user by mistake, it is

represented by a dashed double-headed arrow. The result is a two-way graph of hyperlinks between keywords and pages.

Among them, assuming that search category C has the best similarity to the search keywords, rankC is 1. RankC is 0.5 if it ranks second, and rankC is 0.25 if it ranks third. SimC is Sim (q, c), and numC is the number of data in the search list. If a search result list has not been processed by search category aggregation, rankC is 0.5 and simC is 0.1. Assume that the search category with the highest relevance to the search keyword has a relevance greater than 0.1. In addition, if the lengths of all lists obtained from the search are the same, the score base of NC1 is greater than the score base of NC. This will cause the data score in NC1 to be higher than the data score in NC. In view of the flaws of the standard Rocchio algorithm from Figure 2, it is assumed that the user input search keyword is new, but there is no such search keyword in the personal search feature matrix and the general search feature matrix. Then, the data that appear in multiple search result lists are higher than the score that appears alone. Firstly, each retrieval result data is scored to obtain the total score, and then the data appearing in the different retrieval result queues are summarized into a list and sorted according to the weight from the largest to the smallest.

Then, the relevance of the search keywords and all search categories is 0; that is, the scoring base $W\text{-}j$ of the search list is 0. At this time, the system will only return the data list in the NC. Using the search category with higher relevance for keyword search, the returned data score is recorded as x, and the search category with lower relevance is used for keyword search, and the obtained data score is recorded as y; then, $x > y$ must be (the parameters rankC and simC in $W\text{-}j$ are used to guarantee this rule). If a search keyword is grouped into the wrong search category, the result data of the search will be very small. When all the data scores are calculated, they are sent to the user in descending order of the scores, and the number of feedback data is recorded as M. Then, in multiple search result lists, there are several data with consistent scores; then, the higher the score base $W\text{-}j$ of the column where the data is located, the higher the ranking of the data.

2.1.2. Retrieval Model Parameter Optimization Processing. Figure 3 shows the structure of the acquisition system, which is mainly composed of the following parts. The acquisition control module is mainly responsible for parsing the system configuration file and controlling the operation of the entire system according to the relevant attributes in the configuration file. The control module is also responsible for the management and data communication between multiple acquisition subthreads in the parallel system. The collection module is responsible for managing multilevel queues and accessing the corresponding web pages according to them. The link extraction module is mainly responsible for analyzing hyperlinks from the source code of web pages, analyzing the format of the hyperlinks, and analyzing the hostname and requested file name and has the function of judging the weight. Implementation of the protocol analysis module is responsible for requesting files from the corresponding host, determining the inaccessible website directories according to the contents of the files, and feeding the directories back to the collection module.

The nonrepetitive parsing by the link extraction module will be stored in the buffer area. The buffer area is composed of a queue to be captured, a queue for successful capture, and a queue for failed capture. Among them, the queues to be fetched are divided into multilevel queues according to the order of priority. This module is mainly responsible for converting the domain name of the web server into. This module requires multithreaded security features and guarantees the high speed of message communication. The buffer module stores the visited counterparts in the buffer according to a certain strategy to minimize the number of requests. According to a certain strategy, the web page library extracts themes, contents, and information from the downloaded web pages and saves them in the file system. Internet tourism resources are preserved in a uniform format to ensure the efficiency of visits. The web page download module uses the protocol to send and receive the data returned by the server through asynchronous technology. The theme collection module uses the theme collection algorithm to establish a related database and collects the pages related to the theme based on the collection database. The early research of the PageRank algorithm was mainly used in Figure 4 for the sorting problem of search result page sets, and it has been successfully applied to the topic relevance prediction module of search URLs to be sorted. It can be seen that the research on the PageRank algorithm can better determine the topic relevance of web pages to improve the accuracy of the subject-oriented sorting search strategy.

The current mainstream search engine Google in the Internet industry uses the PageRank algorithm. If a search keyword is grouped into the wrong search category, the result data of the search will be very small. When all the data scores are calculated, they are sent to the user in descending order of the scores, and the number of feedback data is recorded. The basic idea of the algorithm is that it calculates the PR value of each web page in the result page set and determines the topic relevance according to the PR value, and thus determines the web page's relevance. If a web page is linked more often, its importance is higher. In this directed graph, the PR value of node q is t.

3. Application and Analysis of the Internet Tourism Resource Retrieval Model Based on PageRank Search Ranking Algorithm

3.1. Retrieval Model Feature Matching. In order to verify that the search performance of the topic ranking search model based on semantic understanding and dynamic web pages proposed in this paper is better than general web search ranking, the following three aspects are tested: (1) Compare the query performance of the keyword query interface and the query interface based on keyword semantic expansion. Select a set of words as the user query keywords, obtain the

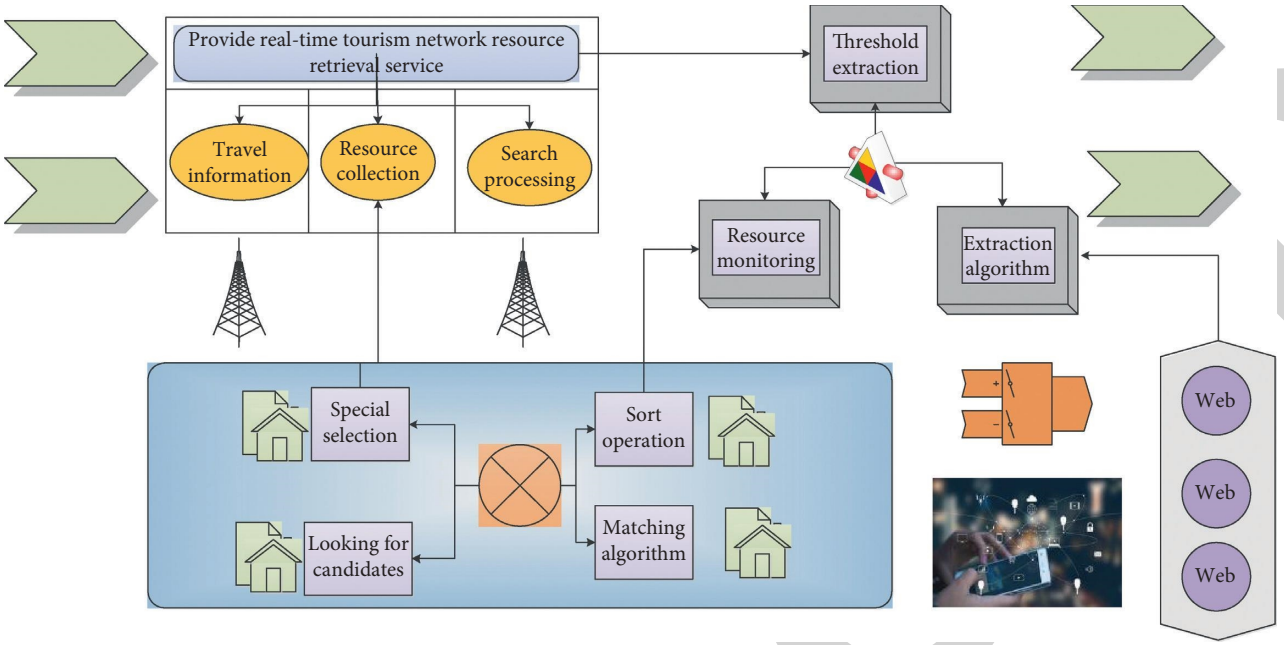

Figure 1: Hierarchical distribution of tourism resources retrieval.

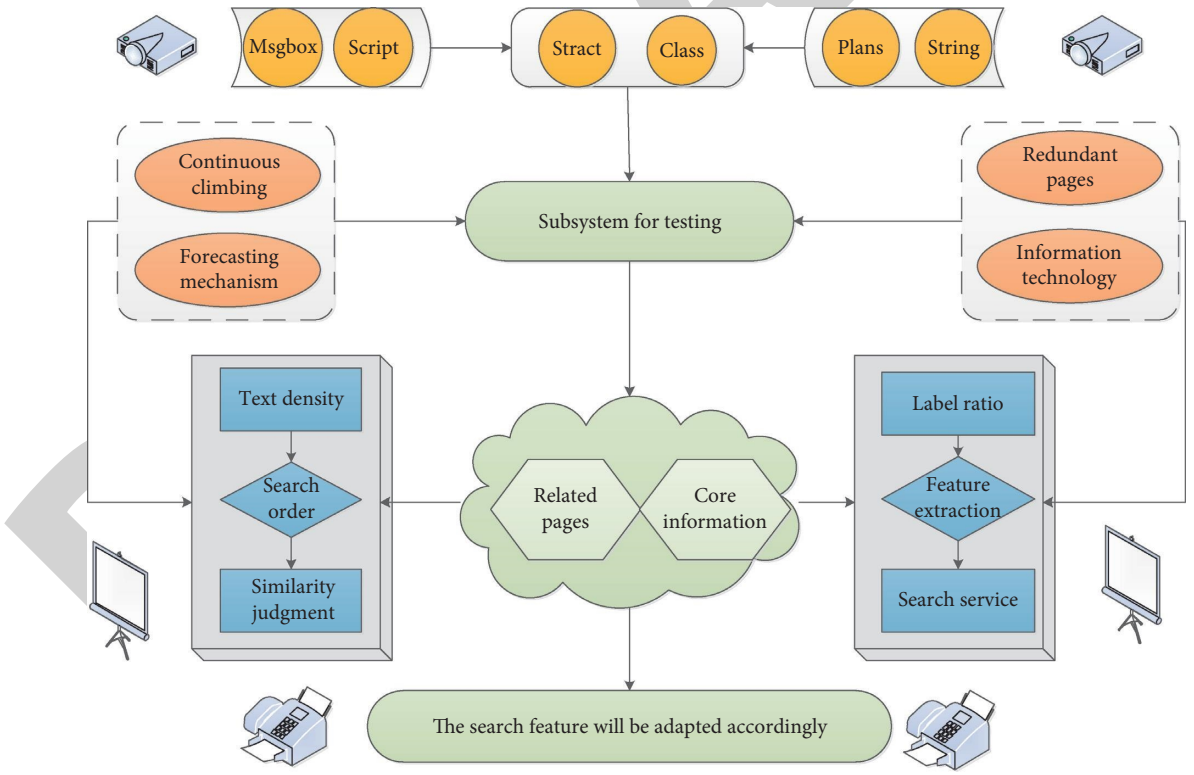

Figure 2: PageRank search ranking algorithm process.

user extended keywords through semantic expansion, and then use Nutch's network search for sorting from the test site. Randomly crawl 5000 web pages, use keywords and extend keywords as user query keywords to call Nutch's full-text search to obtain query results, and compare the query efficiency by comparing the returned result pages. (2) Compare the ranking search performance of static web search ranking and dynamic web search ranking. Select a dynamic website with a travel theme as the test site, and run Nutch web search ranking and dynamic web search ranking

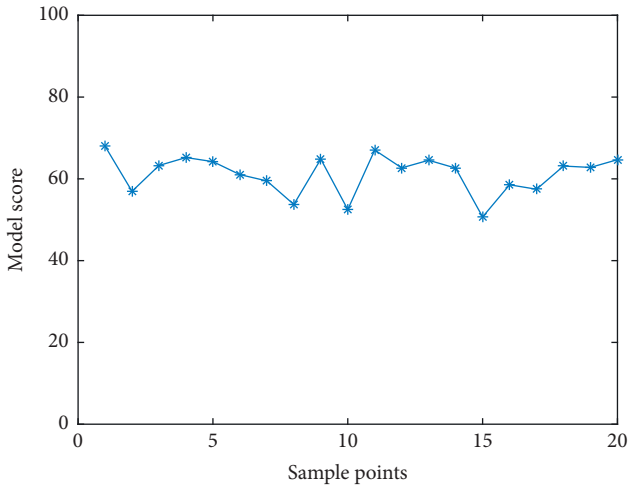

Figure 3: Retrieval model score line chart.

three times under the same software and hardware environment, which lasted 10 h, 15 h, and 20 h. After the search, sorting, and crawling work is finished, all the web pages that it crawls are indexed and the size of the web index file is recorded.

The final result should not include the above three parts, as shown in Figure 5. The Internet tourism resource density algorithm first reads Internet tourism resources by row, counts the number of nonlabel characters in each row and records it as a sample, records the number of characters belonging to the label in each row, and calculates the ratio of the two. What it needs is special attention. The literature algorithm and the algorithm proposed in this paper are used to calculate the user retrieval feature matrix M, and then the user retrieval feature matching algorithm is used for category matching, and the matching accuracy is calculated. The calculated rows are stored in a one-dimensional array in the memory, and then the first-class clustering algorithm is used for clustering so as to extract the content of Internet tourism resources. Before clustering the data in the one-dimensional array, the data need to be smoothed. If data smoothing is not performed, some important data may be lost, such as news headlines. Because these Internet travel resources read by line may be too short, below the threshold of the clustering algorithm, they are discarded by the clustering algorithm.

By giving a specified radius length, calculate the smoothing value of each element in the one-dimensional array. Throughout the experiment, the total number of rows is used for calculation. In order to test the correctness of the algorithm and the results of the clustering algorithm, the experimental results must be compared with the results of manual analysis. For the scientificity and correctness of the test, two test standards are used for the test. The first test method uses the longest common subsequence to calculate the longest common subsequence between manual extraction and extraction. Before calculating, you need to remove special tags, blank lines, and extra spaces. Therefore, it is necessary to provide a relevant start at the beginning, and it is necessary to provide a method for judging the relevance of the page. Topic similarity indicates the similarity between the page and the topic. The pivot calculation is used to judge whether a page is a pivot page and whether it is suitable as the initial similarity in Figure 6. It is used to judge whether the web pages collected by the system are related to a specific topic.

Since the PR value of all web pages is calculated offline, the algorithm has a short response time in practical applications and has good search performance. However, the algorithm does not consider the theme characteristics of the web page. It can be seen from the results in the figure that the average retrieval accuracy of the algorithm proposed in this article is higher than that of the standard algorithm. It does not mean that the page is related to the topic, which will cause the topic of the search result page set to be irrelevant, that is, the phenomenon of topic drift, which not only consumes network resources but also wastes user time. Therefore, the PageRank algorithm for topic search T-PageRank is proposed, which combines the topic relevance of a web page with its PR value to calculate the topic relevance of a given web page. Since the PR value is the probability of a web page being accessed in the physical sense, the initial value can be assumed to be $1/N$, where N is the total number of web pages. In general, the sum of the PR values of all web pages is 1. In addition to linking A to D, A also links C and B, so when the user visits A, there is a possibility of jumping to B, C, or D, and the jumping probability is 1/3.

3.2. Function Realization of Search Sorting Algorithm. In order to verify the effectiveness of the algorithm proposed in this article, a cross-simulation test is performed on it. Divide the user's travel search records into 10 subsets, each with the same number of travel search records. Run the retrieval algorithm 10 times for each different data subset, and use 9 of them as the training set. If the average value is greater than the standard deviation of all data, then the cluster is likely to be a web page body segment.

The experiment selects (scenery, destination, hotel, ticket, and food) as user query keywords and user expansion keywords obtained through semantic expansion. Then, use the original query keywords and the expanded query keywords as the user query terms, and use Nutch's full-text search to query to obtain the query results. From Figure 7, we can see that in the 30,000 web pages randomly crawled by Nutch's network search ranking, the user query keywords are semantically expanded and compared with the original user query keywords and the accuracy has been improved. It can be seen from the results that the accuracy of the three hybrid feature threshold extraction matching algorithms is not much different, but they are all more accurate than the user retrieval feature matching general retrieval feature matching, so the hybrid feature threshold extraction matching algorithm is better than other algorithms. 2. Compare the ranking search performance of static web page search ranking and dynamic web page search ranking. As a test site, run Nutch network search ranking and dynamic web search ranking 5 h, 8 h, and 10 h under the same software and hardware environment, and get the number of web pages searched by ranking.

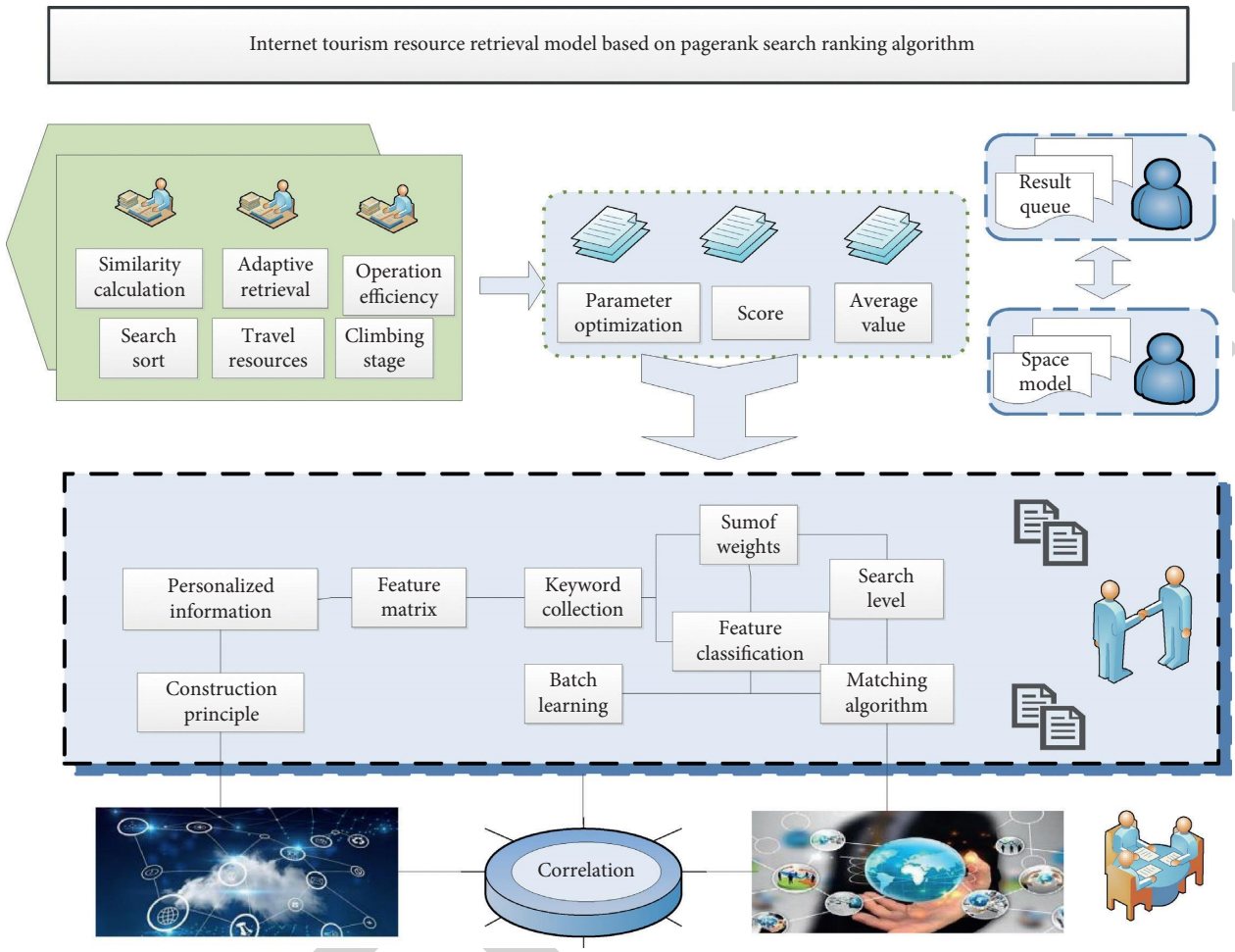

Figure 4: Internet tourism resource retrieval framework using PageRank search ranking algorithm.

Use the Rocchio algorithm proposed in this article to calculate the user retrieval feature matrix M and then use the user retrieval feature matching algorithm for category matching and calculate the matching accuracy. In order to further verify the performance of the algorithm, consider using the mixed feature threshold extraction matching algorithm and gradually increase the training set. The specific accuracy comparison of the three matching algorithms is shown in Figure 8. The above experiment shows that when the data training set is small, the accuracy of the user search feature matching algorithm is lower than that of the general search matching algorithm. Even if the training set is small, the hybrid feature threshold extraction matching algorithm can still obtain better results. When the training set gradually increases, the accuracy of the user retrieval feature matching strategy and the hybrid feature threshold extraction matching strategy will increase.

3.3. Example Results and Analysis. The experimental environment is as follows: hardware environment, 4 GHZ memory, 230 G hard disk, CPU 4-core Intel (R) Xeon (R); operating system: MicrosoR Windows XP Professional SP3; software environment: Nutch. 1.4, Eclipse. 3.5.0. This experiment uses the Nutch web search ranking as the general web search ranking and then considers the search strategy of the model proposed in this article from the three aspects of semantic expansion, dynamic web pages, and topic filtering. The search index of the search strategy is compared with the search performance of the topic search ranking based on semantic understanding and dynamic web pages proposed in this article and the general web search ranking. Nutch is an open-source web search engine based on the Java language. It is mainly divided into two functional blocks: network search sorting and full-text search. The main function of the network search sorting function block is to grab web pages from the web and then provide these web pages. The main function of the full-text search function block is to retrieve relevant web pages from the web pages crawled by the network search sort according to the query keywords and return them as results.

After smoothing, it is found that the cohesion within the paragraphs of the article increases, and the difference between the paragraphs increases. The square difference of the entire one-dimensional array is smaller than that before processing, indicating that smoothing has obtained good results. The standard deviation between the data before smoothing is larger, and the standard deviation after data

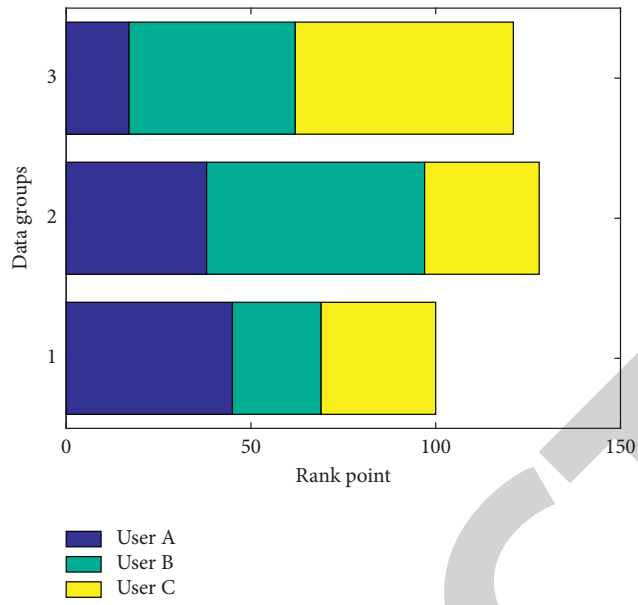

Figure 5: Sorted search stacked histogram distribution map.

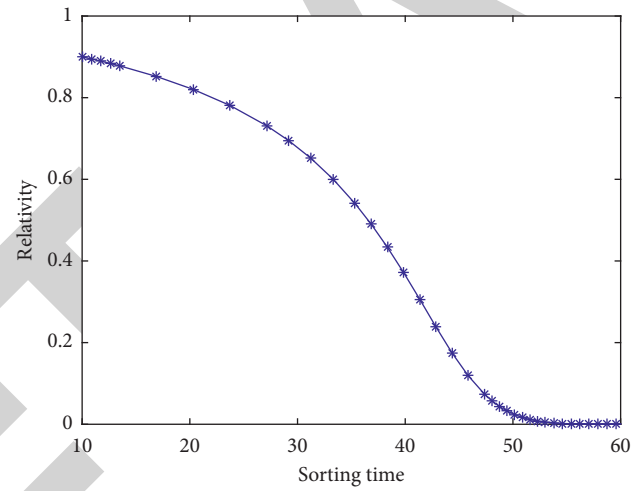

Figure 6: The dependence curve of tourism resource retrieval with the number of searches.

smoothing is reduced. From the perspective of the change in standard deviation in Figure 9, the difference between the data has been further reduced. The above experiment shows that when the data training set is small, the accuracy of the user retrieval feature matching algorithm is lower than that of the general retrieval matching algorithm. In order to verify the effect of data smoothing, two sets of comparative experiments were carried out. In the first group, the clustering operation is performed directly on the group without smoothing processing. The sum of the left and right sides is averaged as the smoothing result. The threshold extraction process calculates the standard deviation of the smoothed array, traverses the array, extracts the rows whose value is greater than the standard deviation, and stores the above-mentioned text rows in the result file.

The above experiments show that when the training set of data is relatively small, the accuracy of the personal feature matching algorithm is lower than that of the general matching algorithm. Even if the training set is small, the hybrid feature matching algorithm can still obtain better results. When the training set gradually increases, the accuracy of the personal feature matching strategy and the hybrid matching strategy will increase. This module mainly includes two processes: the first process is smoothing and the second process is threshold extraction. The highest recall rate can be achieved in this mode, which means that the retrieval effect is the best in this mode. The results of the three strategies are not much different, but all have a certain degree of improvement over the strategy. From the experimental results, the strategy is relatively good. It can be seen

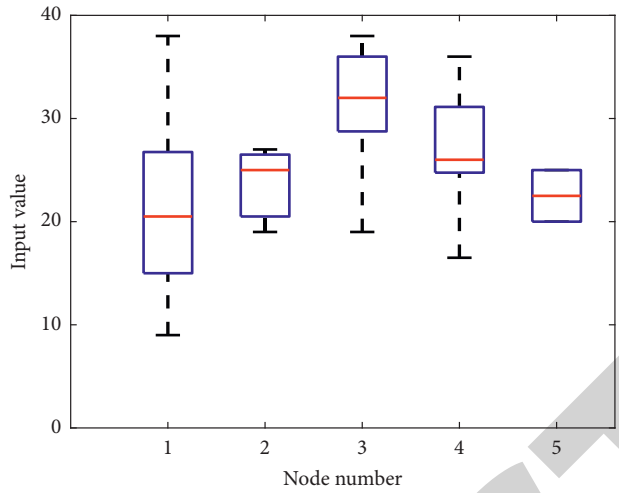

Figure 7: The box plot of the algorithm value changing with the number of nodes.

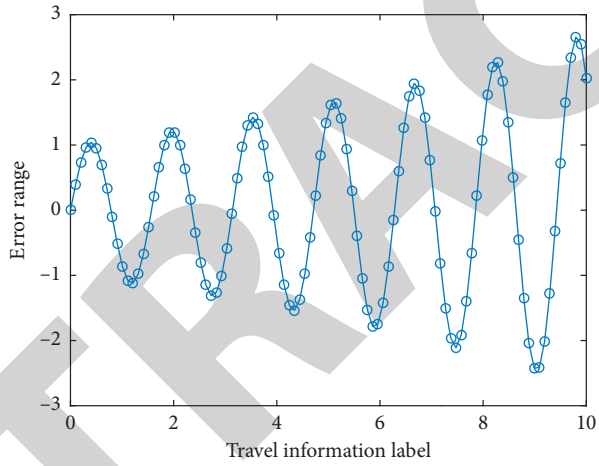

Figure 8: Error curves of different tourism resource levels.

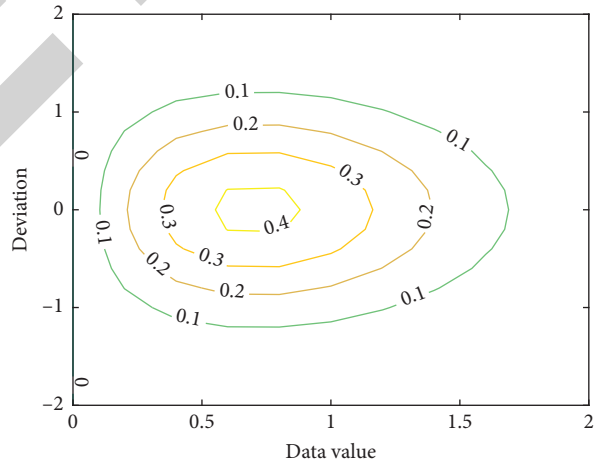

Figure 9: Retrieving model data value deviation contour distribution.

from Figure 10 that the topic ranking search strategy based on the domain topic has a higher precision rate. Through the above tests, we can see that semantic expansion of user query keywords can improve the accuracy of user queries; compared with static web search rankings, dynamic web search rankings are slower to search on designated test sites but

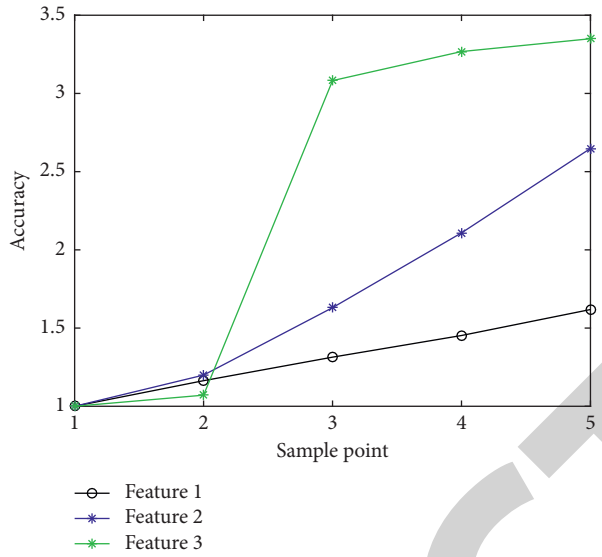

Figure 10: Accuracy line graph of algorithm sample points.

have higher search results. After the construction of the personal search feature matrix M is completed, the correlation between the search categories in the matrix and the search keywords can be obtained, and the correlation score can be performed. The statistical results show its superiority. The topic ranking search model based on semantic understanding and dynamic web pages comprehensively considers the semantic expansion of user query keywords, dynamic web search ranking, and topic filtering strategies and is superior to general web search ranking in terms of recall and accuracy.

4. Conclusion

This article studies and implements the tourism network resource monitoring system and expounds related algorithms and related technologies used in the development of the system. The main work of the thesis is to give the relevant requirements of the theme collection subsystem of travel network resources and then describe the key technologies involved in the theme collection of travel topics such as topic similarity, pivot value calculation, and similarity judgment. The subject collection process is divided into the first sorting search phase, the learning phase, and the continuous sorting search phase. We give an experimental evaluation method and study it through the tourism network resource monitoring system. The experiment verified the performance of the topic capture. An Internet tourism resource extraction algorithm based on Internet tourism resource density is given, and an improved method of data smoothing is proposed. The smoothed data are clustered to extract the main Internet tourism resource content of the web page, give the final experimental results to realize the personalized retrieval subsystem, and use the feature matrix to express the user's interest characteristics and its improved algorithm. Three mixed feature matching strategies are proposed, and the matching effects are compared. An improved score-based web ranking algorithm is used, and a comparison of experimental results is given to realize the tourism network resource monitoring system, introduce the module functions according to the system modules, and give the system screenshots to show the operating effects of the system. The system uses the Internet tourism resource extraction algorithm based on the Internet tourism resource density to remove the noise data from the web pages and improve the response time of the system. Internet tourism resource extraction technology also brings great convenience to data processing.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

The author acknowledges Guangdong Provincial Education Department's Innovative and Strong School Project 2018 (Provincial Key Platform and Major Scientific Research Projects): Guangdong-Hong Kong-Macao Greater Bay Area Marine Tourism Talents Training Research (Project number: 2018GXJK250); and Guangdong Provincial Education Department's 2018 Key Scientific Research Projects in Universities and Colleges: Guangdong-Hong Kong-Macao Greater Bay Area Tourism Industry and Regional Economic Coupling and Coordination Development Research (Project number: 2018WTSCX203).

References

[1] F. Lamberti, A. Sanna, and C. Demartini, "A relation-based page rank algorithm for semantic Web search engines," *Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering*, vol. 21, no. 1, pp. 123–136, 2020.

[2] B. Pan, Z. Xiang, and R. Law, "The dynamics of search engine marketing for tourist destinations," *Journal of Travel Research*, vol. 50, no. 4, pp. 365–377, 2019.

[3] T. Nguyen, D. Camacho, and E. Jung, "Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services," *Personal and Ubiquitous Computing*, vol. 21, no. 2, pp. 267–279, 2019.

[4] Z. Xiang, K. Wöber, and R. Fesenmaier, "Representation of the online tourism domain in search engines," *Journal of Travel Research*, vol. 47, no. 2, pp. 137–150, 2019.

[5] S. Salavati and H. Hashim, "Website adoption and performance by Iranian hotels," *Tourism Management*, vol. 4, no. 6, pp. 367–374, 2018.

[6] T. Agryzkov, L. Oliver, and L. Tortosa, "An algorithm for ranking the nodes of an urban network based on the concept of PageRank vector," *Applied Mathematics and Computation*, vol. 21, no. 9, pp. 2186–2193, 2017.

[7] C. Patel, K. Supekar, and Y. Lee, "OntoKhoj: a semantic Web portal for ontology searching, ranking and classification," *Web Information and Data Management*, vol. 20, no. 3, pp. 58–61, 2019.

[8] M. Gayar, E. Mekky, and A. Atwan, "Enhanced search engine using proposed framework and ranking algorithm based on semantic relations," *Institute of Electrical and Electronics Engineers Access*, vol. 7, no. 3, pp. 139337–139349, 2018.

[9] E. Marine, "A Webometric analysis of travel blogs and review hosting: the case of Catalonia," *Journal of Travel & Tourism Marketing*, vol. 31, no. 3, pp. 381–396, 2019.

[10] F. Gleich and A. Rossi, "A dynamical system for pagerank with time-dependent teleportation," *Internet Mathematics*, vol. 10, no. 2, pp. 188–217, 2018.

[11] F. Chung, A. Tsiatas, and W. Xu, "Dirichlet pagerank and ranking algorithms based on trust and distrust," *Internet Mathematics*, vol. 9, no. 1, pp. 113–134, 2018.

[12] R. Elbarougy, G. Behery, and A. Khatib, "Extractive Arabic text summarization using modified PageRank algorithm," *Egyptian Informatics Journal*, vol. 21, no. 2, pp. 73–81, 2019.

[13] A. Granka, "The politics of search: a decade retrospective," *The Information Society*, vol. 26, no. 5, pp. 364–374, 2019.

[14] J. Valverde and A. Sicilia, "A survey of approaches for ranking on the Web of data," *Information Retrieval*, vol. 17, no. 4, pp. 295–325, 2018.

[15] L. Lechani, M. Boughanem, and M. Daoud, "Evaluation of contextual information retrieval effectiveness: overview of issues and research," *Knowledge and Information Systems*, vol. 24, no. 1, pp. 1–34, 2019.

[16] L. Yang, J. Johnstone, and C. Zhang, "Ranking canonical views for tourist attractions," *Multimedia Tools and Applications*, vol. 46, no. 2, pp. 573–589, 2020.

[17] V. Jain and M. Varma, "Learning to re-rank: query-dependent image re-ranking using click data," *Proceedings of World Wide Web*, vol. 20, no. 11, pp. 277–286, 2019.

[18] E. Marine-Roig and A. Clavé, "Tourism analytics with massive user-generated content: a case study of Barcelona," *Journal of Destination Marketing & Management*, vol. 4, no. 3, pp. 162–172, 2020.

[19] R. Ji, X. Xie, and H. Yao, "Mining city landmarks from blogs by graph modeling," *ACM Multimedia*, vol. 20, no. 9, pp. 105–114, 2020.

[20] G. Tsekouropoulos, Z. Andreopoulou, and C. Koliouska, "Internet functions in marketing: multicriteria ranking of agricultural SMEs Websites in Greece," *Agrárinformatika/ journal of Agricultural Informatics*, vol. 4, no. 2, pp. 22–36, 2019.

[21] R. Dragusin, P. Petcu, and C. Lioma, "Specialized tools are needed when searching the Web for rare disease diagnoses," *Rare Diseases*, vol. 1, no. 1, pp. 528–538, 2020.

[22] K. Leng, R. Kumar, and K. Singh, "Link-based spam algorithms in adversarial information retrieval," *Cybernetics and Systems*, vol. 43, no. 6, pp. 459–475, 2019.

[23] T. Tri and J. Jung, "Exploiting geotagged resources to spatial ranking by extending hits algorithm," *Computer Science and Information Systems*, vol. 12, no. 1, pp. 185–201, 2019.

[24] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: a coauthorship network analysis," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 10, pp. 2107–2118, 2020.

[25] Y. Chen, T. Suel, and A. Markowetz, "Efficient query processing in geographic Web search engines," *Management of Data*, vol. 20, no. 6, pp. 277–288, 2019.

[26] W. Litvin, E. Goldsmith, and B. Pan, "Electronic word-of-mouth in hospitality and tourism management," *Tourism Management*, vol. 29, no. 3, pp. 458–468, 2019.

[27] M. Harb, R. Khalifa, and M. Ishkewy, "Personal search engine based on user interests and modified page rank," *Computer Engineering & Systems*, vol. 20, no. 5, pp. 12–24, 2019.

[28] Z. Xiang, B. Pan, and R. Fesenmaier, "Foundations of search engine marketing for tourist destinations," *The Routledge Handbook of Tourism Marketing*, vol. 2, no. 14, pp. 505–519, 2019.

[29] S. He, F. Guo, and Q. Zou, "MRMD2. 0: a python tool for machine learning with feature ranking and reduction," *Current Bioinformatics*, vol. 15, no. 10, pp. 1213–1221, 2020.

[30] J. Yang, J. Zhang, C. Ma, H. Wang, J. Zhang, and G. Zheng, "Deep learning-based edge caching for multi-cluster heterogeneous networks," *Neural Computing and Applications*, vol. 32, no. 19, pp. 15317–15328, 2020.

[31] Z.-P. Fan, G.-M. Li, and Y. Liu, "Processes and methods of information fusion for ranking products based on online reviews: an overview," *Information Fusion*, vol. 60, pp. 87–97, 2020.

[32] W. Wang, Z. Gong, J. Ren et al., "Venue topic model–enhanced joint graph modelling for citation recommendation in scholarly big data," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–15, 2020.

[33] Y. Li and J. Yang, "Few-shot cotton pest recognition and terminal realization," *Computers and Electronics in Agriculture*, vol. 169, Article ID 105240, 2020.

[34] B. Yang, X. Li, Y. Hou et al., "Non-invasive (non-contact) measurements of human thermal physiology signals and thermal comfort/discomfort poses-A review," *Energy and Buildings*, vol. 2020, Article ID 110261, 2020.

[35] W. Wei, Q. Ke, J. Nowak et al., "Accurate and fast URL phishing detector: a convolutional neural network approach," *Computer Networks*, vol. 178, Article ID 107275, 2020.

[36] J. Wen, J. Yang, B. Jiang et al., "Big data driven marine environment information forecasting: a time series prediction network," *Institute of Electrical and Electronics Engineers Transactions on Fuzzy Systems*, vol. 2, pp. 31–39, 2020.