

PAPER • OPEN ACCESS

Research and Design of Theme Image Crawler Based on Difference Hash Algorithm

To cite this article: De-zhi Wang and Jun-yan Liang 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **563** 042080

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

Research and Design of Theme Image Crawler Based on Difference Hash Algorithm

De-zhi WANG¹, Jun-yan LIANG²

¹Department of computer engineering, North China Institute of Science and Technology, Yanjiao, 065201, China

²Library, North China Institute of Science and Technology, Yanjiao, 065201, China

wangdz@ncist.edu.cn, lijy8011@sohu.com

Abstract. For the problem of high repetition rate of image resources collected by general theme crawler, a theme image crawler system is designed to reduce image similarity. The main contents of the design include the main function modules of the crawler, the workflow of the system and the implementation method of the key modules. The difference hash algorithm is used to solve the problem of image similarity effectively. Combined with Web text cosine correlation algorithm and link PageRank algorithm, the paper comprehensively evaluates the relevance between Web resources and topics. The experimental results show that the subject image crawler can effectively reduce the similarity of the collected images and improve the efficiency of crawler image resources acquisition.

1. Introduction

In recent years, in the research of theme web crawler, it mainly focuses on the analysis of the relationship between keyword information and the importance of web links. Based on the analysis of quantitative data such as the similarity between text content or links and keywords and the number of links, a web crawler model is established to crawl related topic web pages.. However, with the development of the Internet era, the dissemination of information on the network has gradually developed from the traditional simple text dissemination to the direction of multimedia resources dissemination. Among them, information dissemination represented by pictures has become one of the key contents. However, because the Internet is a massive, heterogeneous, dynamic and loosely managed structure, resulting in a large number of identical or similar pictures on the network. This results in a large number of identical or similar images when crawlers are used to collect picture information on the Internet, resulting in a waste of resources. Therefore, how to design a reasonable theme picture web crawler has become a research content.

The main difference between topic image crawler and general topic web crawler is that the focus of general topic crawler is to analyze the text information of web pages. It only uses pictures as a download resource without considering the analysis and utilization of pictures. Baidu, Sogou and other search engines focus on the corresponding logo search of words and pictures. They search the web pages by keywords, and cache the image resources contained in the web pages, instead of deleting the pages without image resources, which results in a large number of identical or similar images in the image search. In the study of image similarity, it mainly focuses on the off-line analysis of images. The corresponding perception model is established by extracting the specific perception information such as the change of gray gradient value, the change of pixel value from spatial domain to temporal



domain, and the change of gray average value. Based on the vector data of the bit eigenvalue model, the correlation of the data is measured, and the image similarity is deduced. Because of its simplicity, rapidity and high accuracy, the difference hashing algorithm is widely used in applications with high time requirement for similarity detection. Therefore, by using image difference hash similarity comparison algorithm and text keyword correlation analysis method, we design a theme web crawler for pictures to achieve efficient and accurate theme web image crawling.

2. Theme image crawler architecture

Theme image crawler is mainly composed of three dynamic libraries and five processing modules. Keyword list repository stores keywords related to topics set manually in advance. The URL queue libraries to be retrieved are used to store useful link addresses in web pages that are dynamically crawled for analysis. The image hash value library is used to store valid pictures, their difference hash values and text information downloaded by the crawler, and is stored in the form of a dictionary. In the dynamic processing module, the initial URL seed queue is used to store the list of first crawled Web pages that are manually selected. The web page acquisition module retrieves the address from the initial seed queue or the URL queue library to be retrieved, and returns the HTML web page data by requesting the relevant URL address. Web page cleaning module mainly uses regular expressions to analyze the acquired HTML pages, extract valid labels and corresponding text information, as well as the image information of Web pages. Text correlation analysis module mainly completes the calculation of data similarity between the cleaned HTML information and keyword list database. The image similarity analysis module mainly completes the similarity calculation between the image in the web page containing the image after cleaning and the hash value library that has been downloaded. Web page comprehensive correlation evaluation module mainly completes the fusion and quantitative calculation of text correlation analysis and image similarity analysis data, and stores the relevant data into the image hash value database and the URL queue database to be retrieved based on the results. The crawler frame of the specific theme picture is shown in Figure 1.

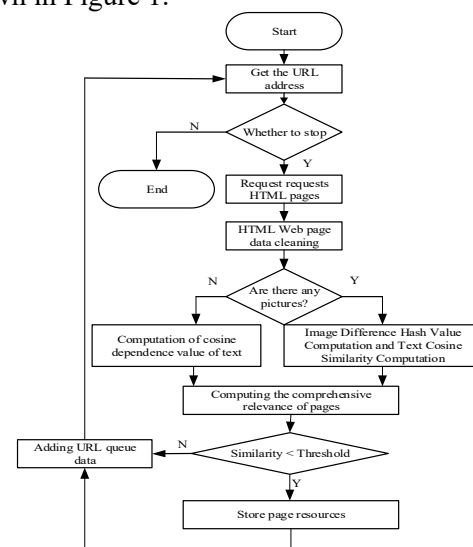
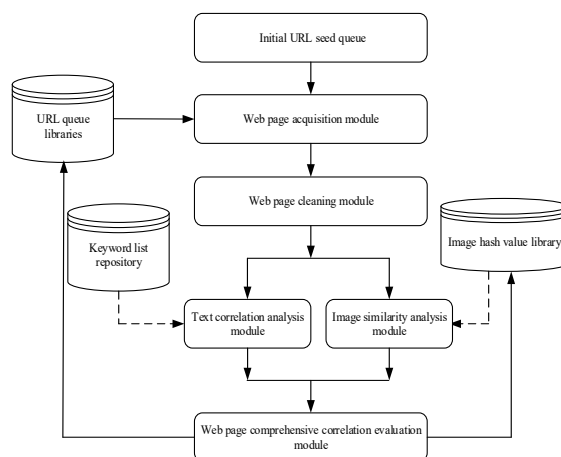


Figure 1. Theme Image Crawler Architecture Figure 2. Theme Image Crawler Workflow Diagram

3. Workflow of theme image crawler

The main work flow of this crawler system is shown in Figure 2. The system first obtains the web page URL address with crawl from the URL queue, and then determines whether the stop condition is satisfied. If not, the HTML page number string that receives the response from the request is used. By extracting the regular expression data of HTML page strings, the valid text data and picture data of the page are obtained. To determine whether the page contains pictures or not, if there are no pictures, use

TF-IDF model to calculate the cosine correlation of text keywords. If there are pictures, the difference hash algorithm is used to calculate the similarity of pictures on the basis of cosine correlation. Combining text correlation and image similarity, the hybrid correlation of web pages is calculated. If the mixed correlation is greater than the threshold, the image resources of the page are stored, the page links are analyzed, and URLs are added to the queue library. If the similarity is less than the threshold, only the URL analysis is performed and the URL queue library data is added. The crawler repeatedly retrieves the address from the URL queue library for analysis until it stops working under certain conditions.

4. Key technologies of theme image crawler

In the design of theme image crawler, the key technologies are image similarity calculation and text keyword correlation calculation, as well as the calculation method of comprehensive page similarity based on this. The efficiency and stability of the crawler can only be guaranteed by constructing a reasonable and efficient similarity technique.

4.1. Computation of difference Hash value similarity for Web page picture

Computing the difference hash similarity of the image is mainly based on the change of color gradient between adjacent pixels in the image, which is processed by binary serialization to generate the corresponding hash value string. The similarity between the two pictures is based on the Hamming distance of the difference hash value string corresponding to the picture. The smaller the Hamming distance, the more similar the two pictures are. Finding the similarity between two pictures mainly includes the following steps.

(1) Reduction of picture standardization

Because the image downloaded by the crawler has different resolution, if only the original image is used for calculation, it will produce different bits of hash value, which is not conducive to the calculation of Hamming distance. Moreover, the calculation takes too many resources and takes too long. It is not conducive to the efficient operation of reptiles. Therefore, first of all, the image is standardized to reduce to h row, l column, that is, the new image is $P = h * l$ pixels. To generate a 64 bit difference hash, reduce the image to $8 * 8$ size.

(2) Grayscale image

Since there are three **RGB** channels in the color picture, and each color channel has a range of $[0, 255]$, the gray value of each pixel can be calculated according to formula (1).

$$\begin{aligned} \text{Gray}(i, j) &= k_1 * R(i, j) + k_2 * G(i, j) + k_3 * B(i, j) \\ 1 &= k_1 + k_2 + k_3 \end{aligned} \quad (1)$$

In the above formula, i and j represent the horizontal and vertical coordinate values of each pixel. $R(i, j)$, $G(i, j)$, $B(i, j)$ are the color channel values of each pixel (i, j) . The k_1 , k_2 and k_3 are the weight values of each color in the gray level, and their sum is 1. Typical values are 0.299, 0.587 and 0.144.

(3) Calculating the difference between adjacent nodes

Since the image has been gray standardized, each pixel $x(i, j)$ has only one gray value $\text{Gray}(i, j)$, a picture can be represented by a data vector $G = (g_1, g_2, g_3, \dots, g_n)$. Among them, g_k represents the gray value of $x(i, j)$ of the pixels. The range of k is $k \in [0, 1]$. The $n = h * l$ is the total number of pixels after image reduction. According to the following formula (2), the gray value difference of each pixel g_k is calculated.

$$\begin{aligned} d(k) &= \begin{cases} 1 & g_k - g_{k+1} \geq 0 \\ 0 & g_k - g_{k+1} < 0 \end{cases} \\ g_k &= \begin{cases} g_k & (k+1) < h * l \\ 0 & (k+1) \geq h * l \end{cases} \\ k &= (i-1) * l + j \quad 1 \leq i \leq h \quad 1 \leq j \leq l \end{aligned} \quad (2)$$

The $d(k)$ in the formula represents the trend of gray level contrast between the pixel $x(i, j)$ and its adjacent nodes. The gray level decreases to 1, and the gray level increases to 0. The h and l represent the number of rows and columns of the picture, respectively. After calculation, a new difference hash binary data vector $D = (d_1, d_2, d_3, \dots, d_n)$ is obtained.

(4) Similarity value meter

According to formula (2), a difference hash binary data vector $D_k = (d_1, d_2, d_3, \dots, d_n)$ is calculated for each picture. The comparison of similarity between the two pictures is expressed by Hamming distance. The cumulative Hamming distance of the two pictures is calculated by summing up the Hamming distance. The smaller the cumulative Hamming distance is, the more similar the two pictures are. The similarity of two pictures is expressed by decimal numbers, and the calculation formula is shown in formula (3).

$$p(m, n) = \frac{h * l - \sum_{k=1}^{h * l} w_k}{h * l} \quad (3)$$

$$w_k = \begin{cases} 1 & |d_k^m - d_k^n| \neq 0 \\ 0 & |d_k^m - d_k^n| = 0 \end{cases}$$

The $p(m, n)$ denotes the similarity between two pictures, ranging from $p(m, n) \in [0, 1]$. The larger the value, the more similar it is. The w_k is the distance between m and n difference hash binary data vectors of two pictures at k position. If the value d_k is the same, the distance is 0 and the distance is different, then the distance is 1.

Through the analysis of the image similarity calculation method based on difference hash value, we can see that this method can extract the trend similarity of the gray level change of each image, but has nothing to do with the specific gray value, so it can effectively deal with the color change of the image comparison collected in the network. At the same time, because of the standardized reduction of the image in the initial processing of the image, it can also effectively deal with the image comparison with different resolutions collected in the network. The disadvantage of this method is that, because the difference hash value is calculated according to the position of adjacent nodes, the Hamming distance increases greatly and the image contrast similarity decreases when the image rotation or screenshot changes greatly. Through a large number of experimental data analysis, in this case, the difference hash binary data vector has similar binary number combination, but not in the same location. Therefore, the distance between Ming and Han Dynasty can be calculated by cyclic moving difference hash binary data vector, and the maximum value can be taken as the final similarity calculation, so as to improve the accuracy of image similarity.

4.2. Computation of cosine coherence degree of page text

Firstly, the weight value $W = (w_1, w_2, w_3, \dots, w_n)$ of the n keywords in the keyword vector space is designed according to the empirical value. Then, using TF-IDF algorithm, the weight value $T = (t_1, t_2, t_3, \dots, t_n)$ of Web text in keyword vector space is calculated according to formula (4).

$$t_i = \frac{\text{count}_i}{\sum_k \text{count}_k} \times \lg\left(\frac{N}{df_i}\right) \quad 1 < i < n \quad (4)$$

In the above formula, the count_i denotes the number of occurrences of word i in the text of a web page. The word frequency value TF is obtained by dividing the number of occurrences of word i in the text by the total number of word segmentation in all web pages. In the formula, n is the sum of all the analysis pages and df_i is the number of articles with word i . Reverse file frequency IDF is obtained by calculating the logarithm of two data quotients. By multiplying the TF and IDF values, the weight of word i in the keyword space vector is obtained.

According to the keyword vector key weight value $W = (w_1, w_2, w_3, \dots, w_n)$ and the web text, the keyword weight value $T = (t_1, t_2, t_3, \dots, t_n)$ is calculated. The cosine value of the vector angle between the

two vectors in the keyword tool can be calculated to get the correlation between the web text and the keywords. The calculation is shown in formula (5).

$$p_k = \cos(\theta) = \frac{\sum_{i=1}^n (w_i \times t_i)}{\sqrt{\sum_{i=1}^n (w_i)^2} \times \sqrt{\sum_{i=1}^n (t_i)^2}} \quad (5)$$

In the formula, p_k denotes the correlation between the k page and the subject keywords in the web page set $P=\{p_1, p_2, p_3, \dots, p_m\}$. The range of p_k is $p_k \in [0, 1]$, and the larger the value, the higher the relevance of the web page text.

4.3. Page comprehensive relevance calculation

Because this crawler system collects thematic pictures, it is necessary to make a comprehensive evaluation of the text information and image similarity of the web pages. Web pages with the same text content or page pictures or similarities are deleted, so as to avoid duplicate collection of resource content and improve the quality of collected data. The comprehensive evaluation calculation adopted is shown in formula (6).

$$P_{mix} = w_1 \times p_k + w_2 \times p_{(m,n)} \quad (6)$$

$$1 = w_1 + w_2 \quad w_1, w_2 \in [0, 1]$$

In the formula, w_1 and w_2 are the weight values of text correlation and image similarity. The more similar the pictures are, the higher the similarity of the web resources will be, thus reducing the value of the comprehensive relevance of the pages. On the contrary, there is no correlation between pictures. The more relevant the text is, the more valuable the image resources are.

4.4. Evaluation and computation of external link URLs of Web pages

This crawler adopts a breadth search strategy, so in the process of crawler expansion, it is necessary to update the URL queue value continuously. The PageRank algorithm is used to calculate the value of the URL for the link addresses contained in the page. When the value is greater than a specific threshold, it is added to the URL queue, waiting for the crawler to collect web resources. Formula (7) is used to evaluate URLs.

$$PR_u = (1-a) + a \sum_{v \in B_N} \frac{PR_v}{N_v} \quad a \in [0, 1] \quad (7)$$

The PR_v in the formula represents the PR value of page v , and there is a link to page u in page v . The N_v represents the sum of all links in page v . a denotes the damping coefficient, which is generally 0.85. The PageRank value of a page is related to the number and quality of the URL pages it joins. The higher the correlation, the higher the PR value of the pages connected. At the same time, the number of links out of the page is inversely proportional to the PR value of the links in the page. The larger the number of links out, the smaller the contribution to the PR value of the links in the page.

5. Experimental results and analysis

Based on the above methods, the experiment of crawler data acquisition was carried out with the theme of "disaster". At the same time, compared with the general topic crawler without image similarity comparison algorithm, it collects and compares the image resources. Figure 3 shows a comparison of the time spent by two reptiles crawling different numbers of pictures. Figure 4 shows a comparison of similar images crawled from the same number of images.

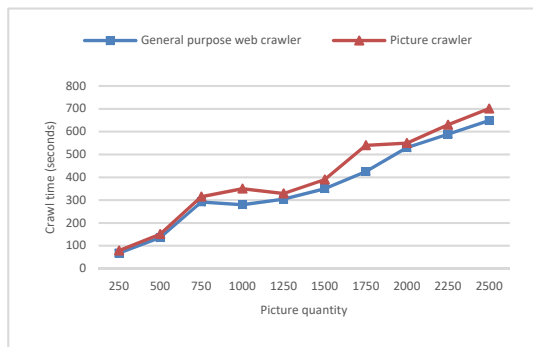


Figure 3. Comparison of crawl times

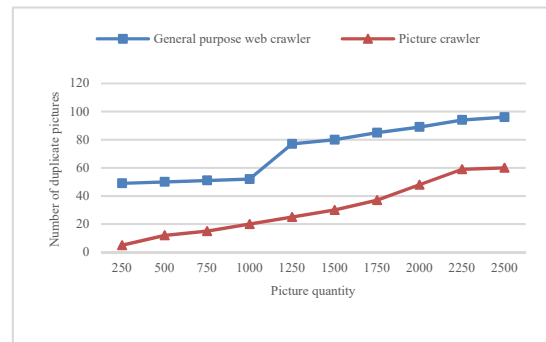


Figure 4. Number comparison of similar pictures

Analysis of Figure 3 shows that under the same conditions, the use time of picture theme crawler and general theme crawler increases, especially with the increase of the number of pictures, the use time is also increasing. This is due to the delay of image similarity comparison. With the increase of comparison library, the time of each comparison is increasing. Analysis of Figure 4 shows that the number of repeated pictures in the picture theme crawler is significantly lower than that in the general theme crawler. But there are still some similar pictures, because some pictures are processed by rotation, alteration or irregular interception, which results in the reduction of similarity. Comprehensive analysis shows that the subject image crawler of this system can obtain higher non-repetitive subject picture resources at the expense of relatively less time.

6. Summary

Subject image crawler based on difference hashing algorithm can effectively improve the collection rate of non-similar images and improve the accuracy of image search. However, in terms of image similarity processing time, the speed of image comparison decreases with the increase of image database. We can improve the similarity comparison algorithm by adding a suitable fast search algorithm. Therefore, further improving the speed of image comparison is the main work to be done in the next step.

Acknowledgement

This research was financially supported by the National Key R&D Program of China[ProjectNo.2018YFC0808306] and the Fundamental Research Funds for the Central Universities of China[ProjectNo.3142015107,3142015107].

References

- [1] Zhang Jin, Ni Xiaojun. Research on topic crawling strategy based on semantic tree and VSM [J]. Computer Technology and Development, 2017, 27 (11): 66-70.
- [2] Zhang Lizhen, Zeng Qingtao, Li Yeli, et al. Research on crawling algorithm for book theme [J]. Journal of Computer Science, 2017, 44 (b11): 460-463.
- [3] Wang Aihua. Design and implementation of vertical search platform for electronic product information [C]// Proc of International Conference on Robots & Intelligent System. Washington DC: IEEE Computer Society, 2017: 101-104.
- [4] Singh S P, Bhatnagar G. A robust image hashing based on discrete wavelet transform[C]//Proceedings of the 2017 IEEE International Conference on Signal and Image Pro-processing Applications, Kuching, Sep 12-14, 2017. Piscataway: IEEE, 2017:440-444.
- [5] Russell B C, Torralba A, Murphy K P, et al. LabelMe: A Database and Web-Based Tool for Image Annotation[J]. IJCV, 2008, 77(1-3): 157-173.
- [6] Du Y J, Liu W J, Lv X J, et al. An improved focused crawler based on semantic similarity vector space model [J]. Applied Soft Computing, 2015, 36(C): 392-407.

- [7] MO-JI WEI,YAN-QING ZHAO,SHI-WEI ZHU,AI-QIN YANG. The Method of Keyword Based Crawler Load Balancing[P]. DEStech Transactions on Computer Science and Engineering,2018.
- [8] Manish Kumar,Ankit Bindal,Robin Gautam,Rajesh Bhatia. Keyword query based focused Web crawler[J]. Procedia Computer Science,2018,125.
- [9] Xiaojun Liu,Wei Hu. Attention and sentiment of Chinese public toward green buildings based on Sina Weibo[J]. Sustainable Cities and Society,2019,44.
- [10] Hyo-Jung Oh,Dong-Hyun Won,Chonghyuck Kim,Sung-Hee Park,Yong Kim. Design and implementation of crawling algorithm to collect deep web information for web archiving[J]. Data Technologies and Applications,2018,52(2).
- [11] Khoulood Boukadi,Mouna Rekik,Molka Rekik,Hanène Ben-Abdallah. FC4CD: a new SOA-based Focused Crawler for Cloud service Discovery[J]. Computing,2018,100(10).