American Journal of Software Engineering and Applications
2014; 3(6): 90-94
Published online December 22, 2014 (http://www.sciencepublishinggroup.com/j/ajsea)
doi: 10.11648/j.ajsea.20140306.14
ISSN: 2327-2473 (Print); ISSN: 2327-249X (Online)

Design and implementation of image search algorithm

Zhengxi Wei, Pan Zhao, Liren Zhang

School of Computer Science, Sichuan University of Science & Engineering, Zigong Sichuan 643000, PR China

Email address:
413789256@qq.com (Zhengxi Wei), zhaopan@suse.edu.cn (Pan Zhao), zhangliren@whu.edu.cn (Liren Zhang)

To cite this article:
Zhengxi Wei, Pan Zhao, Liren Zhang. Design and Implementation of Image Search Algorithm. *American Journal of Software Engineering and Applications.* Vol. 3, No. 6, 2014, pp. 90-94. doi: 10.11648/j.ajsea.20140306.14

Abstract: Image search is becoming an urgent problem of the next generation of search engine. We firstly review the developed situation of image search engine in this paper. Then, the main difficulty and key technologies about this engine are analyzed. Next, the design method is elaborated in detail, which mainly includes image recognition, perceptual hash algorithm, system solution, image retrieval procedure as well as software module, and so on. As a result, we develop an image search engine according to above design methods and implement searching image on the Internet. The testing results finally prove the overall performance of our image search engine is excellent and achieves the desired design requirements. By using data filtering technology and perceptual hash algorithm, the search time-consumed is less than 1 second and is of high search efficiency.

Keywords: Image Search Engine, Perceptual Hash Algorithm, Image Recognition, Feature Index, Grey Classification

1. Introduction

With the explosive growth of the Internet, Web Search technology marked by keywords has acquired a great success in the tremendous information retrieval. As the network develops into the Web2.0 era, people no longer satisfy with merely the text-search, also want to be able to find more images from the sample image. In the future, image search engine [1] will become the main tool of the user to retrieve images in the network.

Google, Yahoo and Baidu, as three common search engines for users to search kinds of information, have launched keyword-search-image service, but the application of the image-search-image is still in testing phase. The image content is more abundant and more complexity than the text content and its amount of information is also that the text cannot be compared with. In addition, the text itself is able to express some semantic meaning, but images can only be expressed through their own content features. Therefore, image retrieval to be implemented is much more difficult than text retrieval.

Image search software is being developed towards the trend of the intelligence and the diversification. On the one hand, image-processing technology plays a key role to support image retrieval. On the other hand, people have developed many convenient development toolkits, which are capable of establishing image feature database. That makes it possible that the image search technology becomes more and more mature. As the same time, the efficiency of retrieving image becomes higher than that of the past and people's workloads are reduced greatly.

2. Image Recognition and Retrieval

Image search is not available without image recognition [2] and retrieval technology. Design and implementation of image search is based on the latter methods.

2.1. Image Recognition

In general, image recognition process can be divided into three main parts: (1) image preprocessing; (2) image segmentation and extraction feature; (3) the judgment or classification. The block diagram is shown in figure 1.

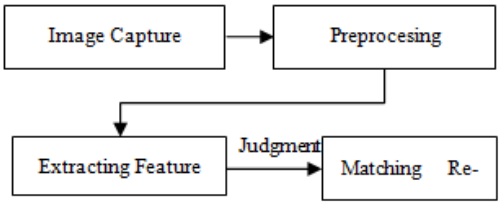

Figure 1. *Image recognition.*

Any kind of image recognition methods first through a variety of sensors convert a variety of physical variables to values or set of symbols that the computer can receive. Traditionally, the space of this value or symbol is called the

pattern space. In order to extract effective identified information from these numbers or symbols, it must be the following processing, eliminating noise, excluding irrelevant signals, calculating feature (such as the shape of the object, perimeter, area, etc.) as well as the necessary transformation (such as Fourier transformation).

Then by feature selection and extraction, the pattern feature-space is established. The subsequent pattern classification or pattern matching is based on the feature space. Finally, the system will output object's the type or a model number that means this object in model database is the most correlative to the object to be searched.

2.2. Image Retrieval

Images can be manually labeled character information based on contents in the text annotations. Web Crawler [3] can collect pictures from the web environment or extract some image marked similar text information in an HTML page, and establish originally keywords. Then it performs a preprocesses to these image, which includes de-nosing, setting standard size etc.. The image is stored to the memory in development board and its feature-index will be further perfected after processing of relevant algorithms. And this index can later be retrieved and compared to the search keywords. In this way, it can determine whether they are the retrieved objects.

The image information will be abstracted to a generic string main through caliphemir algorithm library. Such as color histogram and other information can be extracted through the adjustment of parameters. Open source tools package (Java caliphemir) extracts the features such as color histogram and layout, convert them to the corresponding string from the image. The correspondence between extracted strings and images is established through the inverted algorithm used to file-search, co-exist in the index file. Different picture information can be stored in different fields, together constitute one document for the query. The feature string of image acts as a search keyword, and a picture of the maximum likelihood is found by querying the index file. Finally, a group of image in the picture library is found. Subsequently these pictures are extracted according to their path information.

3. Image Processing Algorithm

The key content of our design is image match algorithm, which is to search the similar images from a sample image according to image feature index. This index is established mainly by image grey value. Grey match can determinate the similarity between two pictures resorting to some measures, for example, correlation function, covariance function, mean square error, etc.. Perceptual hash algorithm [4] is one of the most representative algorithms. The image processing is as follows according to this algorithm.

(1) Formating Image

The resolution of sample image is firstly shrinked to n×n ($n \leq 8$), total n^2 pixels. This can exclude the details of pictures, only leaving some basic information such as structure and gray, and get rid of the differences that the contribute of resolution and brightness gives rise to.

(2) Grey Degree Deduction

Image grey is uniformly decreased to T-class, that is, to make the all pixels be only T kinds of colors, $T \leq 64$ class. Such treatments are designed to exclude the differences of the color number in the image and set the gray value of every pixel into a same range.

(3) Calculate Mean Grey Value of Image

Mean grey value u is figured out in an image composed of total n^2 pixels, using (1). In equation, x_i denotes a value of a pixel and p_i denotes the probability value that this pixel appears in the image.

$$u = \sum x_i \cdot p_i \qquad x_i \in [0, T] \tag{1}$$

(4) Pixel Binary Mapping

According to (2), every pixel is mapped into a binary number. The gray of each pixel is compared with the average value of image, if greater or equal, 1 is recorded; on the contrary, 0 is recorded.

$$f(x_i) = \begin{cases} 1 & x_i > u \\ 0 & x_i \leq u \end{cases} \tag{2}$$

(5) Construct hash value sequences

The comparison result of previous step is combined together, obtaining a sequence of binary integer with n^2 bits, such as {1,0,0,1, ..., 0, 1}. It is exactly the image fingerprint [5] of each picture.

(6) Image Comparison.

Comparison algorithm is to find how many different binary bits between two images with n^2 bits binary integer number, which is equivalent to compute the Hamming distance between images. In general, if the different data bits are not more than five, it shows two pictures are very similar; however, if more than 10, they are two different pictures.

Having carried out the previous five steps of the algorithm, we can calculate image fingerprint from the sample picture, and write the key information including image fingerprint, sample picture location in storage as well as its URL into the database. The next work is to compare the fingerprints of different image and then figures out the similarity degree between the pictures.

4. Software Solution

4.1. Main Function Analysis

Compared with the text search engine, image search engine should complete the following four tasks: collecting images on the Internet, calculating the image similarity, maintaining the image-index library, and responding user's query. So an image search engines need to be provided with the following four basic functions.

(1) Multi-threading Technique

Start multiple threads to surf the Internet and obtain a large of images and their URL. There are two image sources,

namely the image directly from HTML pages or indirectly from the image database on the Web.

(2) Image Feature Extraction

Various image properties (such as the color histogram, the shape histogram, multi-resolution texture features, and so on) are calculated in order to obtain the feature vector and index vector.

(3) DBMS Maintenance

The index records in the database needs to be periodically maintained and updated, and keep the databases time validity and data integrity.

(4) Man-machine Interface

User-oriented image-retrieval interface allows the user can use the example query, such as uploading images, and the interface returns the corresponding result set.

4.2. Software Module

Web crawler (also known as web spider, web robot) is a kind of in accordance with certain rules, automatic program or script to crawl the World Wide Web information. Heritrix is a web crawler developed exclusively for downloading the Internet Webpage. The image search engine uses the Heritrix to download the specified website picture, then the downloaded image will be preprocessed to create image index, and finally the index information will be stored in a Berkeley DB database (a high-performance embedded database). When the user upload pictures to query, the search engine opens images index, matches the corresponding index, and extracts the similar images in the database and return them to the user. The whole process is as following.

1) Capture image
2) Preprocess
3) Extract Feature
4) Judgment and Matching

According to the above model, image search engine is mainly composed of three modules: crawler module, create index module and search module.

(1) Crawler module

Web-crawler Heritrix is used to visit some specific websites and download the pictures in WebPages crawler module: Web crawler Heritrix is used to visit some specific websites and download the pictures in the WebPages. Then, these images will be stored in various hierarchical directories of the local file system.

(2) Create index module

The image information will be abstracted into a generic string main through caliphemir algorithm. Such as color histogram and other information can be extracted through the adjustment of parameters. Open source tools package (Java caliphemir) extracts the features such as color histogram and layout, convert them to the corresponding string from the image.

The correspondence between extracted the string and the image is established through the inverted algorithm that is used to file-search, co-exist in the index file. Different picture information can be stored in different fields, together constitute one document for the query.

(3) Search module

Search module receives the pictures uploaded by users and finishes the flow process: firstly preprocessing user picture, analyzing image feature, next extracting features to compare to the images on the server, obtaining the index number of similar pictures, and then further getting more information(such as image path, index text, etc.) from the image library, at last returning the most similar images to the client for the user to view.

According to the different bit number of image fingerprint, image search engine can calculate the matching degree between the sample image and the similar image.

4.3. Improved Technique and Flow Chart

(1) Data filtering

Heritrix [6] embedded extractor cannot crawl links according to a format, but crawl down all the information in WebPages. As a result, it is not able to specify the content crawled down. This will result in the mirror information is too complicated and there are a lot of redundant information. Obviously, that is not what the search engine needs.

In the extension process to the Heritrix extractor, we use the regular expressions to match the link being crawled down. The matched link will be added to the Heritrix processing queue, only to crawl a specific link (such as the Computer Science Department website, http://jkx.suse.edu.cn/ *) This makes the Heritrix to only crawl the specific data format, play the role of data filtering, and also quicken the speed of data acquisition.

(2) Search flow chart

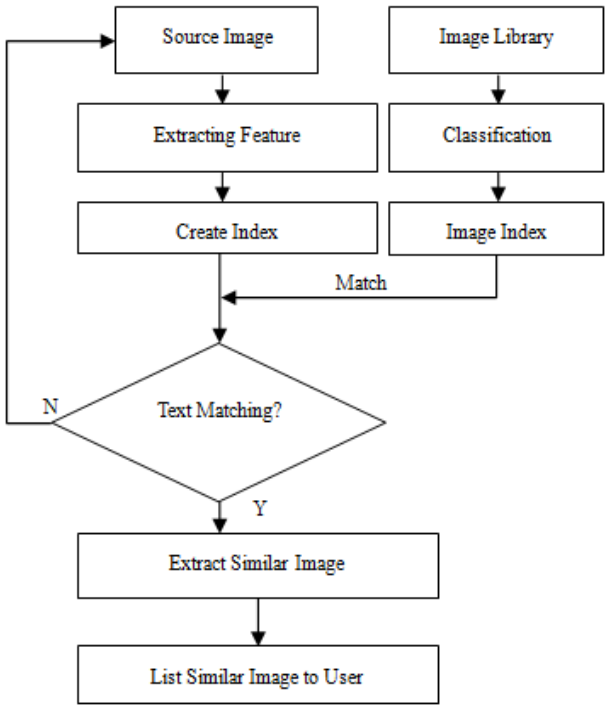

Figure 2. *Search flow chart.*

The search process is shown in Figure 2. The feature string of image acts as a search keyword, and a picture of the

maximum likelihood is found by querying the index file. The most similar image in the picture library is obtained from its path information. The contradistinction process among images is according to the index keywords to match, calculates the image match-degree, and returns the high match-degree images to the user.

(3) Heritrix software toolkit

In the case of default, Heritrix software toolkit uses Hostname-Queue-Assignment-Policy to make that the same host name URL will be placed in a queue. In this way, it can cause that a length of the queue is very long when crawling a single website. According to Heritrix rules, a thread always gets a URL link from the head of the queue. Then the queue will be blocked and it will not recover from the blocking state until that link is processed over.

In order to avoid the occurrence of that case, we adopt the perceptual hash algorithm to solve this problem. This algorithm inherits Queue-Assignment-Policy algorithm and extends its function. It allows the crawling process can create multiple threads so that the entire process will not be suspended due to a single thread blocking. On the other hand, this way also improves the speed of crawling data.

5. Experimental Results

The user interface of our image search engine is shown in Figure 3 and the user clicks on the "Browse" button to select images to be searched. If the selected picture is valid and meets the format requirement, the progress bar will show the name and image size of the picture. The "Upload" button becomes available and the search engine automatically works for users to search similar images.

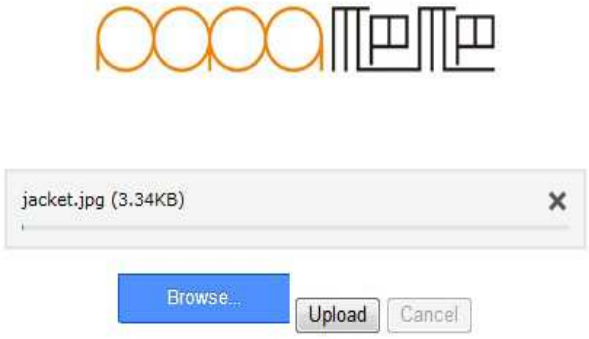

Figure 3. *Upload image interface.*

The image retrieval results are shown in Figure 4. There are some pieces of image and hyperlink addresses in the search result page for users to show URL of the picture. Users are able to visit the relative website as long as they click on these pictures or their URL. Below the URL, we also give the matching degree of the returned images.

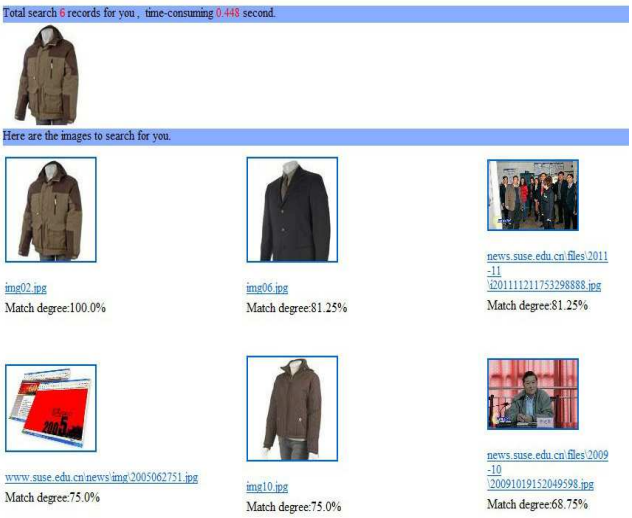

Figure 4. *Search results.*

About 50000 pictures in the image library are used for the test. Even if the amount of pictures is up to 50000, the search engine returns the search results in less than 1 second. The tests show that under this query condition the retrieval speed of the image search engine is also very high efficient.

6. Conclusions

Combined with image-recognition technology, the paper presents a design method for image search engine, which mainly includes image recognition, perceptual hash algorithm, system solution, image retrieval procedure as well as software module, and so on. The experimental results show that the software based on image search algorithm works stably and its search velocity is very fast, compared to the other similar soft wares.

The design idea for developing image search engine can be referenced because the perceptual hash algorithm is well fit to image processing and the method of image grey classification is easy to implement. After a few modifications, our search engine can be applied to more equipments, which is used in not only the personal computer or web workstation, but also mobile phones and other portable devices.

Acknowledgements

The research was supported by Artificial Intelligence Key Laboratory of Sichuan Province (No. 2013RYY04) and the Sichuan Provincial Education Department’s Key Project (No.14ZA0210).

Our work was also supported by university Key Laboratory of Sichuan Province (No. 2013WYY09) and Fund Project of Sichuan Provincial Academician (Experts) Workstation (No.2014YSGZZ02).

References

[1] Cao Y, Wang H, Wang C, et al. Mindfinder: interactive sketch-based image search on millions of images[C]//Proceedings of the international conference on Multimedia. ACM, 2010: 1605-1608.

[2] Zhu B B, Yan J, Li Q, et al. Attacks and design of image recognition CAPTCHAs[C]//Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010: 187-200.

[3] De Groc C. Babouk: Focused web crawling for corpus compilation and automatic terminology extraction[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on. IEEE, 2011, 1: 497-498.

[4] Sarohi H K, Khan F U. Image Retrieval using Perceptual Hashing [J]. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN, 2013: 2278-0661.

[5] Ghose T, Erlikhman G, Garrigan P, et al. Perception, Image Processing and Fingerprint-Matching Expertise[C] //PERCEPTION. 207 BRONDESBURY PARK, LONDON NW2 5JN, ENGLAND: PION LTD, 2013, 42: 11-12.

[6] Liu D F, Fan X S. Study and Application of Web Crawler Algorithm Based on Heritrix [J]. Advanced Materials Research, 2011, 219: 1069-1072.