The Face Database Development of Science and Technology Experts Based on Web Mining

Chunxia Yang
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, China
957539223@qq.com

Yunyang Yan and Quanyin Zhu*
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, China
hyitzqy@126.com

Abstract—For the Face database system by Web mining of science and technology experts, the Web mining data include the faces of science and technology experts, name, gender, title, schools and other properties. The Python programming language and the establishment of the Django framework are used in the proposed system. According to the Web extracting, the basic information of science and technology experts is extracted by the URL filtering. The face data of science and technology experts extracting is obtained by OpenCV libraries. As a result, it can detect face and cut them out. The extraction also adopts the semantic analysis method which uses predefined keywords segmented in the webpage text to calculate the shortest distance. The face date of science and technical experts downloaded are deposited in local folder, and then the absolute path got by means of folder traversal is stored in the MySQL database system which can achieve the query and statistics of basic information of science and technical experts.

Keywords—web mining, science and technology experts, face database system, Python language, face detection, MySQL database

I. INTRODUCTION

As science and technology are developing rapidly, along with the development of computer and Internet technology, more and more affairs are transmitted and processed through the Internet, as a result, web-pages become the important information carrier. At present, most talent management institutions in our country basically have their own construction of expert database, the first group of experts for main storage contains: academicians of the Chinese Academy of Sciences and Chinese Academy of Engineering, the national outstanding professional and technical talents, the national and the municipal expert with significant contribution, experts enjoying special government subsidies, persons selected for New Century Talents Project, personnel with senior title, the municipal outstanding young intellectuals, personnel chosen by municipal science and technology star project and principals of institute, association and council belonging to city association for science and technology[1]. Information management system of six-tech periodicals reviewing experts in Journal of Shanghai University based on the analytic hierarchy process adopts B/S (Browser/Server) structure and the experts' information are recorded so that they can be edited, selected and contacted. The information includes name, gender, date of birth, education degree, professional title, social position, the research direction, work unit, the mailing address, post code, office telephone, home telephone, mobile phone, E-mail address, information-providing unit, upload time, review and censorship situation, etc [2]. From the development experience of the developed countries abroad, we can conclude that all modern enterprises have placed the development and management of human resources at core locations of enterprises management. Although reasonable allocation of funds, technology and material is of great importance, the modern enterprise management will become bosh without the reasonable allocation of people. Development and management of human resource is the organization guarantee to boom enterprises. Personnel information system based on web data mining with open structure indicates that internationally the ability to manage talent resource information has been considered as one of the most important symbol of whether a government can be efficient to perform its functions or not, and is also the future direction of technology progress and application.

II. FACE DATABASE SYSTEM BY WEB MINING OF SCIENCE AND TECHNOLOGY EXPERTS

The system already exists in the faces of the 2990 science and technology experts and the basic information of science and technical experts. The Web mining data include the faces of science and technology experts, name, gender, title, schools and other properties. The system structure of science and technology experts is shown in Fig. 1.

III. WEB MINING

Combining traditional data mining with Web, Web data mining technology can extract potential useful pattern and implicit information automatically and intelligently from mass data in web. Web data mining process: locating resources; getting information from the target in web; information selection and pretreatment (removing obvious errors and redundant data and converting the data to effective forms to make data mining more efficient); pattern discovery (choosing appropriate algorithm to mine new pattern in terms of the mission requirements and expressing it in an understandable form); pattern analysis (verifying and explaining the pattern of former step, and returning to some
steps in the previous processing for reiterative extraction when necessary). Finally, the knowledge will be provided for users in an understandable way [3]. The flow chart of web data mining is shown in Fig. 2.

![Flow chart of web data mining](image)

Figure 1. The system structure of science and technology experts

**IV. WEB CRAWLERS BY PYTHON LANGUAGE**

Web crawler is a kind of program or script automatically to grab the World Wide Web information according to certain rules. It downloads web-pages from the World Wide Web for search engine, and is an important component of the search engine. Beginning from one or some initial web URL, traditional crawler obtains initial web URL. In the process of grabbing web pages, new URLs constantly are extracted in a queue from the current page, until condition in which the system must be stopped is met. Work process of focused crawler is relatively complex, links irrelevant with theme need to be filtered out depending on a certain web analysis algorithm, and useful links waiting for being grabbed are kept and put into URL queue then, it will select the next web URL to grab according to certain search strategies from the queue. The process will be repeated and won’t stop until a certain system condition is met.

Web crawler is written by the system, and a database url is set up in the database in which two tables are set, the top_url table and the table child_url. The first link of URL needing deep excavation is stored in top_url, and URL deeply excavated in child_url. There are many pages hidden deep in the invisible network.

Usually these pages can only be visited when a query is submitted to database, and ordinary crawlers have no access to these pages without links to their words. An attempt is made by Google site map agreement to allow finding these deep resources. Deep page-crawlers increase the number of links for grabbing web-pages. The page link counts. According to the requirement, many links not only take no significance, but also increase rate of mining error, consequently, filter function is used in the crawler code. To improve the efficiency, parallel crawler needs to be written which consists of multiple crawlers running in parallel. Its objective is to maximize the download speed, and try to reduce the parallel cost and repeated download pages at the same time. In order to avoid downloading a page twice, system needs strategy to deal with new-found URL, because the same URL address may be caught by different crawler process. After running the crawler program, a lot information is meaningless, so the efficiency of crawler is quite low. Further research is mainly in how to improve the efficiency of crawlers.

**V. FACE DETECTION BY THE PYTHON LANGUAGE**

Face detection, refers to the judgment whether there is one face in input face image or video, further provision of position, size of each face, position information of each main face organs if there is one, and further extraction of face photo in each image according to these information. Face detection process: first extraction need to detect image, the system is based on analytical URL their web extraction picture of the URL download images, after face detection module, the first picture pretreatment, such as noise reduction, according to the characteristics of the face skin color face detection, if is photos would face to face positioning, pretreatment such as uniform size, and calls the normalization, finally feature extraction, will face region box out, the actual box out of the face region less than the face region, so in cutting face, need will mark the coordinates of the face box to expand, achieve to extract out the whole face, face detection is will appear problem, such as in the whole image of master the proportion of small, how to enlarge photos first, so as to improve the accuracy of face detection, and it is also the brightness of the photo, and side face, etc
can cause certain to detect errors, so in the test for the pictures before pretreatment very important [4-7].

Then, it is necessary to turn the color image into gray images, and we will continue to make a series of processing course, such as image equalization, image enhancement and image filtering, to enhance the important trait information and eliminate the useless information of facial image. The facial image pre-processing steps can be described in Fig. 3.

![Image](image.png)

**Figure 3.** The steps of image preprocessing in face recognition

Python language is based on OpenCV call library implements face detection; OpenCV face detection is mainly calls the trained cascade (Haar classifier) to pattern matching. CvHaarDetectObjects, it will be the first image grey, according to the parameters are introduced into judgment canny edge processing (default don't use), and then to match. Collect the matching of the match out piece, filtering noise, if more than the number of calculation adjacent e. (of incoming min_neighbors) just as output, or cut.

Matching cycle: will match classifier amplification scale (to value) times, at the same time, the narrow scale times, for matching the match the size of the classifier until more than thumper, and return the match results. Matches cvRunHaarClassifierCascade to call for matching, will all the results in CvSeq*Seq (dynamic growth elements sequence), will the results to cvHaarDetectObjects.

CvRunHaarClassifierCascade function is introduced into the whole image and according to cascade to match, and according to the different type to cascade, tree, stump (incomplete tree) or other) for different matching method.

CvRunHaarClassifierCascade function for singles to picture of the test. In the call function it first uses cvSetImagesForHaarClassifierCascade before setting integral figure and the right scale coefficients (=> the window size). When the analysis of all rectangular frame by cascading classifier of each layer when return (this is a possible target), or return zero or negative.

Face detection program mainly completes three parts functions, namely loading classifier, and loading test and measurement for image and labeled. This program OpenCV use provided during the "haarcascade_frontalface_alt.xml" file storage target detection classification, loaded with cvLoad function, after forced type conversion. In the OpenCV provide for testing the image of target function is cvHaarDetectObjects, this function uses the pointer to one target objects (such as face) training cascade classifier in the image of the target object contains found in rectangular area, and will these areas as a series of rectangular box to return. Classifier after using the need to be explicit release, the function is used cvReleaseHaarClassifierCascade.

Save the photographs of technical experts downloaded under the local folders, first, create a new folder, in which detected faces and extraction of the technical experts face are stored. When detecting Face, the absolute path of folder needs to be inputted, select a photo, and judge whether there is one face, and provide position, size of each face, position information of each main face organs if there is one, then further extract face photo in each image according to these information. The cycle repeats, and then the face photos of technical experts can be obtained and output in the new folder.

Call OpenCV library to realize the face detection, set the image path needing detecting, and then store and output the folder path of faces. The original figure and rendering are as shown in Fig. 4.

![Image](image.png)

**Figure 4.** The original figure and rendering

VI. EXTRACTING TECHNOLOGICAL EXPERTS' FOUNDAMENTAL INFORMATION

Semantic analysis is a logical stage of the compilation process [8], whose task is to take a examination of related properties of context for the correct procedures structurally and of the types. Semantic analysis reviews the source program to confirm if it has the semantic error, collecting information of types for code generation stage. For instance, one work of semantic analysis is review of the type to see whether each operator have the operand the language standard allow. When it does not accord with the language standard, the compiling program should report the error. For example, some compiling programs inform of the condition that real number is used as the array subscript another example, some programs set that the operand can be forced. Then, when two mesh computing is put on a integer and a full mold object, the compiling program should convert the integer to the full mold and can't regard it is the source program's errors. The process of matching the word by using semantic analysis realizing the segmentation is shown in Fig. 5.
VII. CONCLUSION AND FUTURE WORK

Construction of technical experts face image database based on web data mining technology has wider search range, so mining difficulty is bigger, useful information mined is less. The further improvement is mainly in web crawler, the first is the response time of the mining, the second is further design of filter used in mining, and the last is improvement of accuracy of technical expert’s basic information.

Of course, this is just a preliminary study of our project. There are still many problems to be solved on the face database system by web mining of science and technology experts. Firstly, the samples of each expert are distinct and limited, obviously the increment of samples will improve the detection rate. Secondly, because the quality of every static image is varied, image pre-processing becomes rather important aspect of face filtering. How to improve the quality of image to raise the face detection rate of Web images becomes a huge task in the future work.

ACKNOWLEDGMENT

This work is supported by the National Sparking Plan Project of China (2011GA690190), the fund of Huaian Industry Science and Technology. China (HAG2011052, HAG2011045, HAG2010066).

REFERENCES


