Journal of Computer and Communications, 2017, 5, 60-68
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Automated Picture Crawling for Electrical Power Maintenance

Bing He[1], Xuqi Zhang[2], Mei Wang[1], Xiu Cao[2], Zikun Zhuang[2]

[1]Shanghai Inspection & Maintenance of Power Company, Shanghai, China
[2]School of Computer Science Fudan University, Shanghai, China
Email: hebing_qh@163.com, 15210240105@fudan.edu.cn, meiwang70@hotmail.com, xiucao@fudan.edu.cn, zikunzhuang@fudan.edu.cn

How to cite this paper: He, B., Zhang, X.Q., Wang, M., Cao, X. and Zhuang, Z.K. (2017) Automated Picture Crawling for Electrical Power Maintenance. *Journal of Computer and Communications*, **5**, 60-68.
https://doi.org/10.4236/jcc.2017.53007

Received: February 27, 2017
Accepted: March 10, 2017
Published: March 13, 2017

Abstract

Defect recognition based on picture analysis is one of the most important means to detect key failure points or damages. However, the recognition rate is low due to the limited number of pictures collected on site, so the computer training set is limited, which leads to a lack of studying and training. Meanwhile, the internet provides a large number of related pictures which can be used as an important data source for training picture analyzing engines. By using an internet spider under certain rules, one can freely collect information on the internet. The internet spider described in this paper can automatically collect related images from the internet as well as search for similar images by leveraging on a local seed picture. This spider also has a parallel version, which can give significant performance boost when run.

Keywords

Electrical Power Maintenance, Web Crawler, Image Search, Parallel

1. Introduction

Electricity is China's important resource as well as the basis for industrial development. To ensure that the power network runs safely, we are researching focusing on how to do safety inspections for our power network. The means of inspections include manual inspection, robot inspection, manned helicopter inspection, as well as UAV inspection [1] [2]. The traditional way of manual inspection requires inspection personnel to know relevant professional knowledge, and is also very dependent on the work attitude of these personnel. Extreme environment also serves as a great challenge to manual inspection. The other means of inspections, on the other hand, are more efficient, less costly and less risky. Many power line inspection photos will be taken during these modern methods of inspections. Using these inspection photos as a source, it is possible to develop a system that can intelligently analyze the condition of the power network.

During the development of an intelligent analysis system, the size of the image resource set has a decisive influence on the intelligent recognition rate of the system. Currently, there are only a limited number of pictures collected on site, so the computer training set is becomes too limited for enough studying and training, which results in low recognition rates. Meanwhile, the internet provides a large number of related pictures which can be used as an important data source for training picture analyzing engines. This project mainly researches on how to collect pictures related to electrical power maintenance with spider, a way to provide structured training set data for the future development of an image analysis system, so that its recognition can become more efficient as well as more accurate.

2. Related Work

2.1. The Current Situation of Power Maintenance Work

A daily manned inspection is the most widely used method in safety inspections for the power network. However, it is exhausting, time-consuming, and very inefficient. The massive data collected from the inspections are also hard to keep, easy to lose, ill-formed, and difficult to inquire. These data cannot be effectively used for future research. Also, due to the broad distribution of power lines and the geographical limits of manned inspection, many power lines cannot be inspected by workers themselves.

The SCADA (Supervisory Control and Data Acquisition) system is a DCS and automatic electricity monitoring system based on computers. It has a wide application, including data collection, monitoring, as well as process controlling in the fields like electricity, metallurgy, oil production, gas production, chemical industry and railroads [3]. It can monitor as well as control working devices on the spot, so it has many functions including data collecting, device controlling, measurements, parameter adjusting, as well as giving alarms. However, SCADA has its own defects: 1. The monitor terminals have strict requirements for working conditions. 2. The complex and varied communication system does not offer reliable data transmission. 3. The rate of coverage is low while the cost is high.

Helicopter inspection was put into use 20 years ago. This means of inspection has the advantage of quick line patrolling and free of geographical restrictions, but it requires precision flying. The helicopter must fly parallel to the power line, keeping at an almost constant distance with it, including where the power line sags. This gives great pressure to the pilot, both mentally and technically, so non professionally trained pilots can hardly take up the job [4].

2.2. Smart Inspection Based on UAVs

UAV is a new technology that has greatly developed in recent years. It has been used in many fields such as inspection for power lines, water sources, air pollution and plantation. UAVs can also be used in many aspects in electric power, including power line planning, topographic map measuring, power line inspection, disaster response and power line erection [5].

UAVs can carry monitoring equipment to inspect the power lines, and such a method has significant advantage over traditional means of inspection: 1. It is unmanned so it is safer by avoiding casualties. 2 It's free from geographical restrictions, and even when an area is struck by natural disasters such as earthquakes or flood, UAVs can still inspect power lines in these disaster-stricken areas. 3. Quick inspection at over 10 km/h. UAVs do face limitations such as plane weight capacity, weather and airspace approval, but as the UAV technology quickly develops and airspace management becomes more efficient, such limitations will be out of the way in future. Using UAVs on unmanned power line inspections, and forming a system of UAV inspection by integrating advanced technologies from fields like electronics, communication and image recognition provide a quick and safe way of inspection and make up the many disadvantages of manned inspection. Therefore, UAV inspection is of great importance in terms of research and has a great prospect [6].

3. Internet Crawler

3.1. Introduction to the Internet Crawler

The internet is a web-like space of information, and can be described in a directed graph G = (N, E), with contents in a webpage as nodes, marked uniquely by URLs, and links as directed edges, as in **Figure 1**.

The node set is N = {N0, Nm}, and E is the collection of hyperlinks. Leaf nodes can be webpage files, as well as media files like images and sound. All none-leaf nodes are webpage files. When the crawler is crawling webpages, it uses traversal algorithm (depth-first algorithm and breadth-first algorithm) to traverse them.

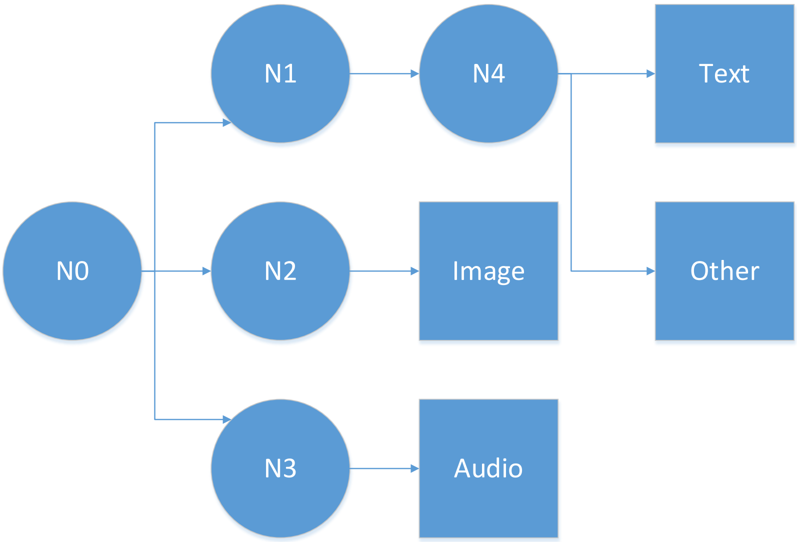

Figure 1. Web structure of the internet.

3.2. The Crawler's Search Tactics

The crawler usually uses these two tactics: depth-first and breath-first.

- Depth-first algorithm tactics means starting from the start page, tracking one link after another, and goes to the next start page to track after processing an entire line.
- Breadth-first algorithm tactics means inserting the links, which are found in the downloaded webpages, directly into the end of the queue of to-be-crawled URLs. In other words, the crawler will crawl all the pages linked to the start page, then select one of the linked pages, and continue to crawl all the pages linked in that website.

3.3. The Running Process of the Crawler

A URL list is needed when the crawler starts to serve as its starting point. Put these URLS into the queue of to-be-crawled URLs, and crawler will start from these URLs. It will then take a URL, resolve the DNS, get the host IP, download the webpage corresponding to the URL, and store it into the downloaded webpage library. Next, put these URL into the queue of crawled URLs, analyze the URLs within it, then the other URLs, put them into the queue of to-be-crawled URLs, and start the next cycle. **Figure 2** describes the entire process.

From the perspective of the crawler, we can divide resources on the internet into the following groups:

- Downloaded and unoutdated webpages.
- Downloaded and outdated webpages: The obtained webpage is actually an image or an archive of contents on the internet. The internet is dynamic, and some of its contents have changed overtime. Therefore, these webpages become outdated.

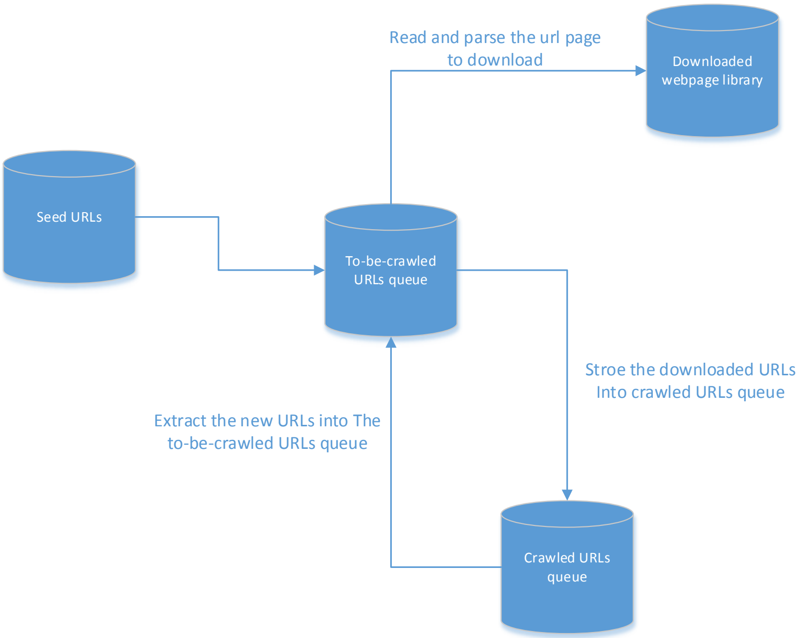

Figure 2. Web crawler URL capture flow chart.

- Webpages to be downloaded: These are the pages in the list of to-be-crawled URLs.
- Knowable webpages: These pages are neither downloaded nor in the list, but they are considered "knowable" since their URLs can be obtained from analyzing downloaded pages and pages corresponding to the URLs in the list.
- Unknowable webpages: These webpages cannot be crawled and downloaded by the crawler directly.

4. The Picture Crawling System Design for Power Maintenance

The internet crawler is an important tool for obtaining online data, and there are countless researches on its technology. Yet, is it difficult for normal internet crawlers to crawl power maintenance images directly from the internet since all major search engines have anti-crawling system that prevent generic crawlers from crawling image resources. To tackle this issue, this paper will talk about making use of the technical principles of the internet crawler and develop a crawler designated towards obtaining power maintenance-related images on the web. After the user types in keywords related to electric power, the system will automatically download images related to the keyword.

4.1. System Framework

The following is the design of the process, which is also illustrated in **Figure 3**:

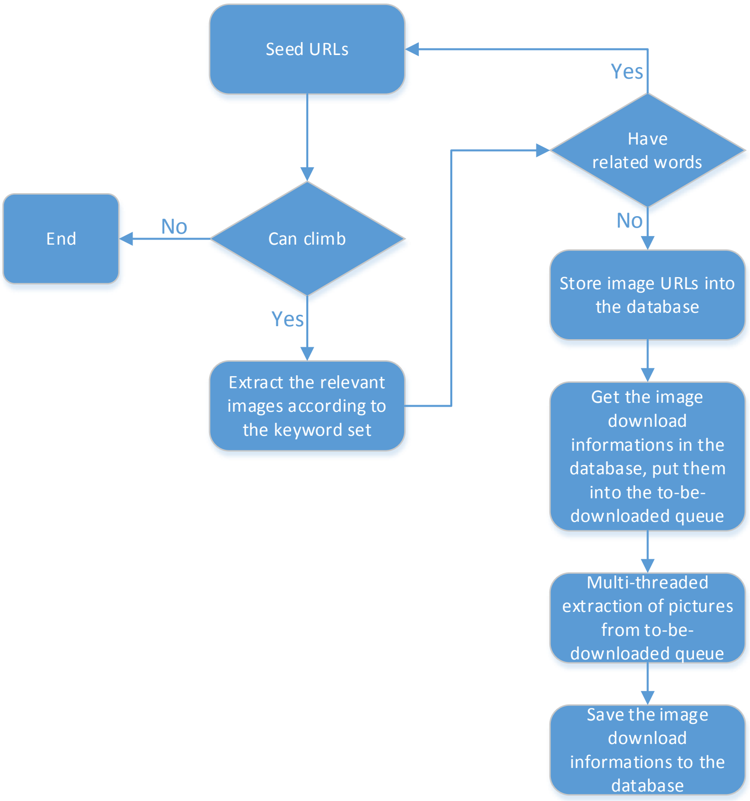
Seed URLs
Yes
Have related words
No
End
No
Can climb
Yes
Store image URLs into the database
Extract the relevant images according to the keyword set
Get the image download informations in the database, put them into the to-be-downloaded queue
Multi-threaded extraction of pictures from to-be-downloaded queue
Save the image download informations to the database

Figure 3. Power image crawler processing flow chart.

1) Obtain the crawl seed (By searching the URLs and the keywords. The URLs and the keywords must be accessible, or the crawler's life cycle will end right here).

2) Extract URL information of images according to the crawl seed (Batch saving pages where the pictures are).

3) Store the saved image URLs into the database (Scanned ones will be set as 1 while unset ones will be set as 0), resolve the URLs, and collect associated seed URL. Go to step 1 if any are found.

4) Thread read unscanned URLs in the databases, and obtain from them download URLs of every image, the store the download URLs into the databases (These URLs will also be set: Unscanned ones as 0, successfully scanned ones as 1, failed to scan ones as 2).

5) Extract unscanned download URLs from the database and add them to the task queue.

6) Download the images, and turn the download process into a persistent operation.

7) Perform step e, or end the process.

4.2. Crawler Queue Design

It is very important to build a generic and scalable crawler queue for large crawler applications. The design of the crawler queue, a queue data structure for saving URLs, is the key to the crawler itself. We designed a fitting queue for the crawler according to the different functions crawler queues have.

- To-be-crawled URL queue and to-be-downloaded image URL queue will use queue as their data structure.
- Crawled URL queue will use Hash as its data structure in order to store massive data while guaranteeing fast writing speed.

4.3. Crawling Images Based on Textual Keywords

The crawler can crawl images from major search engines like "Baidu", "Sogou" and "Bing". This is the process of crawling resources from search engines:

- The client sends keyword search request to the image search servers of search engines.
- The servers receive the request and sends back relevant image download links according to the requested information.
- The client receives the returned links and store them into the database for later download.
- The local downloader downloads undownloaded images by asynchronous multithreading.

These are the issues that need to be solved in the actual experiment:

- The anti-crawling system which is present in many major search engines. The download links, which are returned from the image server after receiving search information, cannot be downloaded directly by programs. However, the original image links can be extracted from the html codes the search engines return by using tools like html parser and regular expression, and such links are available for download to programs. Take Baidu as an example: Baidu returns four image download link properties: "thumbURL", "middleURL", "objURL" and "fromURL", alongside image information returned to the user. Only the original download links obtained from "objURL" property can be downloaded automatically by programs. The links obtained from the other 3 properties will be blocked by Baidu's anti-crawling system when trying to download automatically with a program.

- After obtaining image download URLs from the server, store them into the database. The purpose is to help realizing breakpoint resume for the image downloading task. Also, it will store other relevant information of the pictures, like the download status and the source search engine.
- Correlation downloading for similar keywords. During the experiment, we found that in the crawling process with a single keyword, the relevancy of images will decline as the search goes deeper. In order to collect a large number of relevant pictures while ensuring relevancy, single keyword should be expanded to other similar keywords, as well as the process of crawling under a single keyword to many similar keywords. For example, the single keyword "Electric tower", can be expanded to similar keywords like "Power tower", "Wind power tower", "Wire tower", "High-voltage line" and "High-voltage tower", as illustrated in **Figure 4**.

4.4. Crawling Based on a Seed Image

Search based on textual keywords is now a mature technology. The accuracy of the search depends not only on the amount of webpage information stored in a

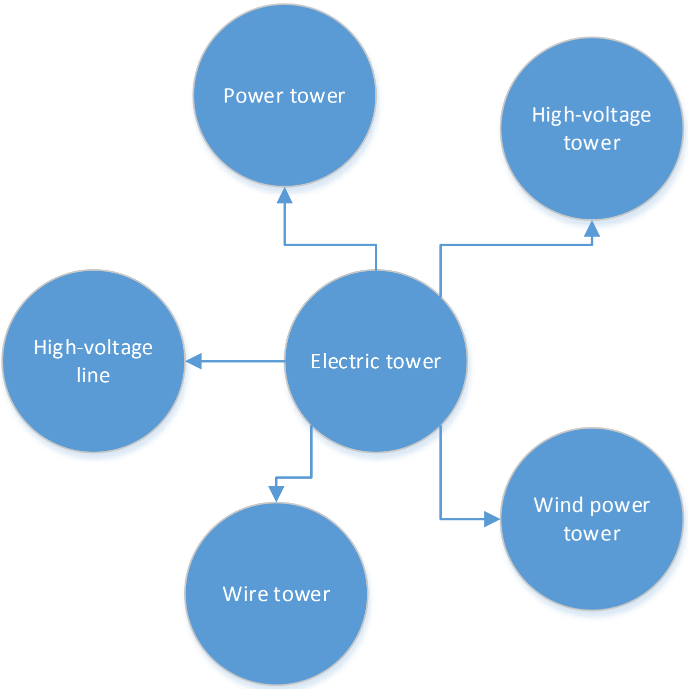

Figure 4. Keyword extension image.

search engine's system and the accuracy of information classifying tags, but also on the target file the user's searching for. If the user is focusing on searching textual information, this type of search gives an accurate feedback. However, traditional text-based search can only analyze textual contents on a webpage and cannot satisfy the growing need for searching online multimedia files, which are ever growing in amount. In terms of searching images, the accuracy will drop evidently because image files in webpages cannot be accurately described by texts. Here we propose a method for image-based searching: By directly analyzing image files based on content search methods, and categorize them by their characteristics. A user may submit an image to show his or her search intentions: For instance, the user can upload a seed image, then the program will analyze and compare the image, and crawl similar images from the contents of the seed image [7].

For instance, this is how to crawl images from an image-based search on Baidu:

- The client uploads a local image to the server through Baidu's Image recognition API.
- Baidu's Image Recognition server analyzes the image's contents and return to the client the contents' keywords.
- The client uses the keywords and use them to crawl images.

4.5. Parallel Crawling

There are a large number of images related to electric power, so it is difficulty to crawl fast with a single-threaded crawler. So the crawler can be expanded to create multi-thread framework for crawling electric power images, in which the crawling is done by a multi-threaded crawler.

The multi-threaded electric power image crawler introduced here uses a master-slave mode, where the master node maintains the entire to-be-crawled URL queue and the to-be-downloaded image URL queue, while the slave node downloads every single image and persists image information into the database. The master node is responsible for obtaining to-be-crawled image information from the internet, and put them into the queue of to-be-crawled images. The slave node takes out one picture's information at a time and crawls the image.

5. Experimental Results

The experiment analyzes the efficiency of the multi-thread crawler designed and demonstrates the result of image crawling with this software.

A single-threaded electric power crawling tool is compared to its multi-threaded versions on downloading 1500 identical images. The processing speeds are listed in **Table 1**.

Table 1. Time comparison of multi-threaded parallel downloading images.

Number of threads	1	3	5	7	9
Time (s)	1430	680	570	472	469

From the data presented in **Table 1**, we can see that the speed dramatically increases when using a multi-threaded crawler than a single-threaded one. As the number of parallels increase, the download speed grows linearly at the beginning, but drops when there are too many threads. The results show that a multi-threaded is more suitable for a massive crawling task, and a proper number of threads must be chosen.

6. Conclusions

As a key technology in the field of computer science, internet crawlers have also great value on using in the electric power industry. This paper introduced a software that automatically crawls electric power related images with the internet crawling technology. This program provides a huge image source for the smart learning of defect recognition based on picture analysis, and will therefore serve as a guide for key inspections and maintenance measures in future smart power line inspections.

A quick electric power crawler has the following main characteristics:

- Easy to use and has a GUI.
- Supports the combination of searching by textual keywords and seed images. The user can customize interested keywords of upload related seed images, and the crawler can automatically match corresponding images and save them to the drive.
- Supports parallel. The parallel function can be used by users who want a large number of images in a short period of time, and this will bring satisfactory results.

References

[1] Han, B. and Shang, F. (2016) A Frame Model of Power Pylon Detection for UAV-based Power Transmission Line Inspection. *Zhejiang Electric Power*, **35**.

[2] Li, H. (2008) Research of Motion Control System for Power Transmission Line Inspecting Robot Based on ARM. North China Electric Power University.

[3] Yu, Y. and Lin, W.-M. (2012) Study on Industrial Control SCADA System's Information Security Protection System. *Netinfo Security*, **5**.

[4] Shen, G.-S. and Zhao, X.-B. (2008) Helicopter Power Line Inspection Tour Technology. *Electric Power Construction*, 29.

[5] Liu, G.-S. and Jia, J.-Q. (2012) UAV Applications and Development in the Power System. *Journal of Northeast China Institute of Electric Power Engineering*, **32**.

[6] Peng, X.-Y., Liu, Z.-J., Mai, X.-M., Luo, Z.-B., Wang, K. and Xie, X.-W. (2015) A Transmission Line Inspection System Based on Remote Sensing:System and Its Key Technologies. *Remote Sensing Information*, **1**.

[7] Wang, C. and Zhao, B.-F. (2012) Content-Based Image Search Engine. *Journal of Changsha University*, **26**.

Scientific Research Publishing

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/
Or contact jcc@scirp.org