Web Crawler Based Study on Traffic Accident Data Acquisition for Operating Tunnels

Zhipeng Liao*, Hao Ding, Yangyuyu Xia and Fei Ma
China Merchants Chongqing Communications Technology Research & Design Institute Co., Ltd., Chongqing, 400067, China
*Corresponding author’s e-mail: 284920980@qq.com

Abstract: With constant development of tunnels, traffic accidents in the tunnels continue to rise. To control and manage tunnels and reduce traffic accidents. Based on Web crawler technology, this paper, based on Web crawler technology, covers research on traffic accident information characteristics for operating tunnels and Web crawler mechanism to establish a PySpider-based Web platform for real-time acquisition of traffic accident data for operating tunnels. This platform can fetch valid traffic accident data URL for operating tunnels from designated websites and to extract, analyze and store traffic accident feature data, thereby managing and controlling traffic risks existing in an operating tunnel, reducing traffic accidents and safeguarding people's properties.

1. Introduction
Recently, with the development of transportation system, by the time of late 2018, there have been 17738 tunnels, with 17236.1 km in length, which makes china the largest number of tunnels, with longest mileage among the world. However, the operation risk of highway tunnels is very high thereupon, with increasingly accidents happened in the tunnels. After the data collection and analysis, the potential risks in operating tunnels can be managed and controlled effectively, in order to decrease the accident incidence. However, the traditional way of data collection is energy and time consuming, not beneficial to mass data collecting and acquisition [1][2].

Regarding traffic accident data for operating tunnels in the context of Internet, this paper discusses how to create a Web crawler based operating tunnel traffic accident database on the basis of fusion and intelligent association of vast amounts of multi-sources and heterogeneous data. In addition, in view of frequent update of different websites on the Internet it presents improved Web crawler design to increase data capture coverage and efficiency.

2. Principle of Web Crawler
Web crawler, also known as Web spider automatic indexer, is a program or an Internet bot that systematically browses the World Wide Web. Web crawler is an important component of the grabbing system of Web search engines [3]. The main purpose of the Web crawler is to download Web pages from the Internet to form an image copy. The Web crawler is widely used by Web search engines and some other sites to update their Web content or indices of others' Web content. Web crawlers can copy all pages visited for processing (sorting downloaded pages) by a search engine which indexes the downloaded pages so users can search more efficiently. To put it simply, they replace your manual operations of opening windows, entering data, and so on with programs which acquire desired information for you and parse and store such information into a database [4]. Fig. 1 illustrates the
basic workflow of a Web crawler:
- Select some URL seeds through access to designated websites;
- Put these URLs into a task queue to be fetched;
- Extract URL from the task queue, parse DNS, obtain the host’s ip, download corresponding Web page and store it into a downloaded Web page library. Meanwhile, put these URLs into the URL queue fetched;
- Analyze URL in the URL queue fetched and other URLs and put URLs into the URL queue to be fetched, thus entering next cycle;
- Parse downloaded Web page and obtain desired data;
- Store the data into database;

Fig 1. Data capture flow of Web crawler
Fig 2. PySpider framework

3. Design and Implementation of Web Crawler
For the purposes of design and implementation of the Web crawler in this paper, Ubuntu16.04 operating system is used as development environment; front end interface is developed using JavaScript, HTML etc; Web crawler at the back end is developed using Python; PySpider is used as Web crawler framework; MySQL5.6 database is used; front end communicates with back end through Http.

3.1. Overall framework
In this paper, PySpider framework is used as the overall design framework for the Web crawler. PySpider is a powerful Web crawler framework with built-in WebUI that allows visual programming and debugging of the crawler. PySpider framework is written using Python language, has a distributed architecture and supports multiple kinds of database as data storage back-end. PySpider supports script editor, task monitor, project manager and result viewer [5]. Its overall architecture is illustrated in Fig. 2.

For each data capture in the PySpider framework, the Scheduler initiates task scheduling, the Fetcher fetches Web content, the Processor parses Web content and then sends generated Web Request to the Scheduler for retry and outputs final crawling result to database for storage.

3.2. Web page parse
The Web crawler crawls the Web page, extracts, organizes and stores useful key information from the Web page [6]. Web page parse techniques that are commonly used include regular expression, BeautifulSoup and PyQuery. PyQuery is jQuery implementation in Python and allows convenient, fast and efficient operation and crawling of Web pages using jQuery grammar. PyQuery can quickly obtain Web page labels to extract key information through CSS selector. Its main code is as follows:
3.3. Design and implementation

The traffic accident data crawler consists mainly of different Web crawler nodes. It has a distributed Web crawler architecture under PySpider framework, and fetches and parses information on operating tunnel traffic accidents through URL queue [7]. The logic design of the distributed Web crawler nodes is illustrated in Fig. 3:

When the crawler starts working, the crawler node uses initial URL entered by the user as initial access and puts it in the task queue. After the crawler node program starts working, the program gets the URL from the task queue for connection and obtains HTML data stream from the URL corresponding Web page. Its main program segment is as follows:

```python
def on_start(self):
    self.crawl(url, callback=self.index_page)
@config(age=10 * 24 * 60 * 60)
def index_page(self, response):
    print(response)
    for each in response.doc('a[href^=url]').items():
        if re.match(url"w+", each.attr.href, re.U):
            self.crawl(each.attr.href, fetch_type='js', js_script=\"function(){
                setTimeout(window.scrollTo(0,document.body.scrollHeight), 5000);
            }\", callback=self.list_page)
```

![Fig 3. Logic design of distributed Web crawler nodes](image-url)
Through PyQuery the crawler parses initial URL Web page, extracts key information and URL links, gets lower-level URLs from index_page, puts them in the task queue, modifies PyQuery parse rules based on lower-level URL Web page features, parses according to lower-level URL Web page features and grabs information on operating tunnel traffic accidents on corresponding Web pages by searching different URLs in a loop. Through constant periodic URL requests, the crawler realizes real-time grabbing of accident information and stores extracted information on operating tunnel accidents into the accident database. The structure of extracted data is complex so JSON is selected as the format of data storage. Extracted data include key information such as accident place, time, type, cause, losses and location [8]. The above design allows the Web crawler to grab information on operating tunnel accidents, effectively fetch data in a more instantaneous and comprehensive manner and improve the efficiency of data capture.

4. Traffic Accident Data Acquisition and Analysis for Operating Tunnels Based on Web Crawler

In order to effectively and correctly acquire internet mass data of accidents in the operating tunnels, Python crawler’ s crawling technology of internet data is employed to acquire the worthy tunnel accident information [9][10]. By the frame of PySpider Python crawler compiles several crawler, and each with specific key words to search and to crawl respective data in the internet, which is analyzed intelligently, and then some of data is extracted and saved in the data [11]. PySpider frame supplies Timer to crawler.by timer, crawler can arouse and carry out crawling task, to fulfil the function of automatic acquisition and extraction of information about operating tunnel accident data. By internet crawler, data about 1165 tunnel accidents at home from 2016 to 2019 are collected. By classifying these data, features of accidents data are analyzed by terms of different time, different type of car, and different causes of accidents, to recognize the potential dangers in operating tunnels, thus supplying the decision foundation in risk management and control [12][13][14].

4.1. Time distribution

By classifying the 1165 tunnel accidents by terms of months, weeks, hours, Fig. 4, Fig. 5 and Fig. 6 are employed accordingly.

![Fig 4. Monthly distribution](image)

![Fig 5. Weekly distribution](image)

![Fig 6. Hour distribution](image)
According to the trend of accidents change, by term of month, tunnel accidents are frequently happened in January, February, April, and October. The causes are January, February are in the peak of spring festival, thus causing the sharp increase of traffic in the tunnel, thereupon high risk of traffic in the tunnel. April, and October are in the peak of tourism, with more tourists on the way to make an increase of traffic load, thereupon high risk of traffic in the tunnel; by term of week, tunnel accidents are happened frequently on Sundays. The cause is there are more people to travel on Sundays. by term of hour, there are an increasing number of accidents from 0 to 8 am, while decreasing number of accidents from 8 to 23 am, and a traffic peak from 8 to 9 am, and a gentle in 10 am. The cause is 8 to 9 am is the peak of working, which makes the sharp increase of traffic in the tunnels, thereupon an increasing traffic risk in the tunnels.

4.2. Vehicle type distribution

In order to study the potential relations between traffic accidents and type of cars, the collected traffic accidents is classified by types of cars, the following Fig. 7 is showed.

![Fig 7. Type of cars](image)

According to Fig. 7, cars and trucks are the main types causing accidents, accounting for 93.1%. the causes are: first, the number of cars and trucks are large among all the type of cars; and second, the car, with small shape and high speed, makes the drivers easy to change road and have an dangerous overspeed. therefore, the incident of accidents is much higher than other types of vehicles.

4.3. Distribution of accident causes

Accident cause is the main cause of traffic accidents in tunnels. Collected traffic data are classified by different cause of accidents in the following Table. 1.

| Classification       | Cause of Accident               | Number of Accidents |
|----------------------|---------------------------------|---------------------|
| Tunnel factor        | Tunnel equipment fault          | 18                  |
|                      | Tunnel lighting insufficiency   | 5                   |
|                      | Icy/slippery road               | 5                   |
| Vehicle factor       | Vehicle failure                 | 53                  |
| Environment factor   | Geologic hazard                 | 19                  |
|                      | Meteorological disaster         | 12                  |
| Human factor         | Illegal overtaking              | 272                 |
|                      | Fatigue driving                 | 346                 |
According to Table 1, human factor is the main factor accounting 92.2%, which is mainly caused by illegal overspeed, fatigue driving, close distance of cars, vehicle overload, and vehicle overspeed, while the vehicle factor accounts for 3.7% and tunnel factors and environment factors account for 4.1%. Based on the analysis above, data support and decision foundation about the recognition of origins of traffic accidents are acquired, thus decreasing incidents of accidents and safe traffic in the tunnels [15].

5. Application of Web Crawler for Tunnel Accident Data
The Web crawler platform for tunnel accident data employs MySQL as its core database. The entire crawler platform runs independently on the server and performs operation, data analysis and storage based on Web platform. Its overall service logic is clear and easy to maintain. Its main interface is shown in Fig. 8.

![Fig 8. Main interface of the platform](image)

The PySpider crawler platform allows daily searching and crawling of several major news websites, sorting of fetched URLs using "tunnel" and "accident" as keywords, putting screened URLs in the task queue for crawling Web page content, parsing fetched Web page data as per accident form, cause, location, time, tunnel type, tunnel length and vehicle type, and storage of parsed accident data in the database. An example of accident information analysis result is provided in Fig. 9.

![Fig 9. Accident information analysis result](image)

The application of the Web crawler based traffic accident data crawler platform for operating tunnels allows efficient acquisition, parse and feature extraction of operating tunnel traffic accident data from vast amounts of big data on the Internet.

6. Conclusions
Based on Web crawler technology, this paper has proposed a traffic accident data acquisition platform for operating tunnels, and developed a tunnel operation accident database system that is able to analyze features of tunnel accident data using Web crawler technology and extract valid accident data.
features. By this, effective data support can be provided for tunnel risk evaluation and management and improve tunnel operation safety can be improved. Tunnel management personnel can prevent accidents in time based on the analysis result of tunnel accident data features, thus significantly improving the tunnel safety and protecting the people's properties.

Acknowledgments
This research was financially supported by the National Key R&D Program of China (2017YFC0806003) and Guangxi Key R & D Plan Project (Guike AB19110019).

References
[1] Ni, H.L., Dai, Y.H., Zhao, Q.X., (2010) Research on distribution characteristics of traffic accidents in expressway tunnel. Highway, 55(4):126-129.
[2] Zhang, Y.C., He, C., Wu, D.X., et al, (2009) Characteristics and countermeasures of traffic accidents in expressway tunnel. Journal of Southwest Jiaotong University, 44(5):776-781.
[3] Ou, Y.C. (2017) Design and Implementation of Distributed Web Crawler System Supporting Dynamic Web Pages Paring, Southeast University.
[4] H, P.F. (2019) Design and Implementation of Cloud Crawler Subsystem of Cloud Data Collecting System, Beijing University of Posts and Telecommunications.
[5] Feng, L., Huang, L., Zeng, L., Y., (2019) Zhu Qihua. Research on web spatial data acquisition based on distributed web crawler. Journal of Guizhou University (Natural Science Edition, 36(01)):33-36.
[6] Zhao, K. (2019) The Design and Implementation of The Topic-focused Web Crawler System, Beijing University of Posts and Telecommunications.
[7] Ding, W.H. (2019) Research and Implementation of Active Defense Technology for Malicious Crawlers, Beijing University of Posts and Telecommunications.
[8] Zeng, J.R., Zhang, Y.G., Zheng, J., Huang, G.J., Chen R.Y., (2019) Implementation technology and application of web crawler for multi-data sources. Computer Science, 46(05):304-309.
[9] Bai, Y.J., Wang, B.F., Xia, Y.Y.Y., (2019) Typical risk identification of Highway tunnel operation. Tunnel Construction, 39(1):471-477.
[10] Lin, Z., Wu, S.Y., (2016) The risk assessment method of tunnel operation safety. China Highway, 5:110-114.
[11] Chen, J.X., Luo, Y.B., Jiang, J.C., (2006) The safety assessment of a road tunnel. modern. TunnellingTechnology, 4:68-71.
[12] occurred in operating highway tunnels in China and their casualties evaluation. Tunnel Construction, 38(4):565-574.
[13] Wang, Q., Jiang, S.P., Liu, H.L., Ma, F., (2019) Analysis of traffic accident characteristics of urban road tunnel. Chinese Journal of Underground Space and Engineering, 15(S1):460-469.
[14] Zhou, P.C., Shi, S.S., (2019) Research on safety technology of highway tunnel operation. China Southern Agricultural Machinery, 50(16):206.
[15] Duan, H.Y., (2019) Current situation and improvement measures of highway tunnel operation safety. Construction & Design for Project, 20:246-248.