Advances in Engineering Research, volume 120
International Forum on Energy, Environment Science and Materials (IFEESM 2017)

Requirements Design for Unidimensional Directional Data Acquisition and Comparative Analysis System——Under Present Situation of Environmental Pollution

Wu Qiong[1, a], Wu Chensheng[2,] Liu Yanjun[3]

[1]No.140, Xizhimenwai Street, Xicheng District, Beijing, 100044 P R CHINA

[a]wwqqppdd@163.com, [b]wu1082@163.com, [c]241743125@qq.com

Key Words: air pollution; system requirements; topical crawler; data

Abstract. In recent years, the issue of environmental pollution has been widely concerned, and a large number of studies have confirmed that the short-term concentration change of air pollution is closely related to the number of deaths per day. As an important cause of death, the respiratory disease is attracting more and more attention. Therefore, this study will propose a requirement design of massive data management system based on web content so as to collect air quality index and traffic index in Beijing City and to provide data support for further modification and establishment of effective environmental laws and regulations.

Introduction

In recent years, due to a series of wide-scale and frequent events from global climate change to regional food security, the relationship between environmental factors and human health is awared of, understood and paid special attentions more and more, the international community, including researchers, policy makers and the public have paid high attention to this problem. With rapid development, China has also suffered serious environmental pollution. People pay the price of health damage while enjoying the achievements of modern society. According to the preliminary estimates of WHO, 22% of the disease burden in China should be attributed to environmental factors. The air pollution is one of the main environmental hazards affecting human health.

A large number of studies have confirmed that the short-term change of air pollution is closely related to the number of human deaths per day. In epidemiological studies, extensive studies have shown that respiratory diseases are most susceptible to air quality. As an important cause of death, respiratory disease is attracting more and more attention. Therefore, this study will propose a requirement design of massive data management system based on web content so as to collect Air Quality Index (AQI) and traffic index in Beijing City and to provide data support for further modification and establishment of effective environmental laws and regulations.

Research Background and Significance

Causes of air pollution

The air pollution is mainly caused by human activities and the air pollution sources include: factory emissions, automobile exhaust, agricultural reclamation, forest fire, cooking smoke (including roadside barbecue) and dust (including construction site). The air pollutants are mainly divided into harmful gases (carbon dioxide, nitrogen oxides, hydrocarbons, photochemical smog and halogen elements) and particulate matter (dust, acid mist and aerosol) that mainly come from factory emissions, automobile exhaust, agricultural reclamation, forest fire, cooking smoke (including roadside barbecue) and dust (including construction site).

The air pollutants are also divided into gaseous pollutants (nitrogen oxides NO_x, sulfur dioxide SO_2, carbon monoxide CO and ozone O_3), persistent organic pollutants (such as dioxins), heavy metals and particulate PM (Particulate Matter).

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Effect of motor vehicles on air quality

The motor vehicle pollution refers to exhaust emission, litter and leak, dust, oil and gas evaporation and engine noise in service, of which the number, concentration and duration exceed the natural purification ability and allowable level of environment, thus causing harm to human living environment. The important pollutants of motor vehicle causing the air quality degradation mainly include particulates ($PM_{2.5}$, PM_{10}), nitrogen oxides (NO, NO_2), carbon monoxide (CO), hydrocarbons (CH), volatile organic compounds (VOCs), ozone (O_3), etc. In addition, the large amount of carbon dioxide emitted from fuel combustion has a long-term impact on the earth's climate system.

In Beijing, great attention has been paid to environmental protection over the years and the air pollution control work has been put in the first place. Since 1998, 16 stages of air pollution control measures have been implemented successively. In 2012, the monitoring and data release of fine particulate matter ($PM_{2.5}$) was firstly carried out in Beijing. With the development of economy, population and energy consumption is growing, and there is a great gap of air quality to meet the people's requirements and the newly-issued ambient air quality standard (GB3095-2012). At present, the air quality in Beijing shows the following characteristics.

Hazards from air pollution

The hazards from air pollution include acute poisoning, chronic poisoning and carcinogenic effects, wherein, the acute poisoning refers to the acute poisoning of people that is not caused by low pollutant concentration of air but caused by massive harmful gas emission and sudden change of outside meteorological conditions under special conditions (e.g., special accident of factory in production process). For example, the methyl isocyanate leak of India Bhopal pesticide factory directly endangered the human body, killed 2500 people and injured about one hundred thousand people; the chronic poisoning indicates chronic toxicity to human health caused by air pollution and mainly presents rising incidence of people after low-concentration and long-time action of pollutants in human body. In recent years, the incidence of lung cancer of city residents of China is high, particularly in Shanghai. The incidence of respiratory disease of city residents was significantly higher than that of suburb residents. The carcinogenic effect is the result of long-term effects by the process that the pollutants stay in the body for a long time, which impairs the genetic material in body and causes mutation, while the body abnormalities of descendants caused by mutation of germ cell is called as teratogenic effect; the caused sudden change effect of organisms genetic material and genetic information is called as mutagenesis; and the caused tumor effect is called as carcinogenesis. The term "cancer" here includes benign and malignant tumors.

In China, the type of air pollution is still the coal smoke pollution, typically SO_2 and PM_{10}. But in many large and medium-sized cities, due to the rapid increase in the number of motor vehicles, the air pollution gradually develops to the combination of coal smoke and automobile exhaust. Compared with developed countries in Europe and America, the air pollution level is quite high in china. According to the 2015 environmental quality bulletin, among 559 surveyed cities in China, there are 210 cities (37.6%) with air quality level of three or below, wherein the average annual concentration of atmospheric PM_{10} in 207 cities is at least 2.1-7.5 times higher than that of European cities; the concentration of SO_2 in 102 cities is 2.8-12.5 times higher than that in European cities. Based on the above background, the air pollution has become one of the important factors affecting people's health and social and economic development.

System Requirement Analysis

Objective of system construction

The unidimensional directional data comparative analysis system is a software system developed by the researchers engaged in the research of air pollution prevention and control for directional data acquisition, intelligent mining and decision analysis. The data directionally

collected by the system includes traffic operation data and air quality data. The data comes from the traffic bulletin issued by Beijing Traffic Information Center and the public data of the Ministry of Environmental Protection of the People's Republic of China. The system performs acquisition, extraction, cleaning, integration and collection of directional data, performs data analysis, mining, processing and storage according to certain rules and criteria, and presents the data in graphical report form. The design purpose of the system is to enable researchers to do a variety of query operations analysis of historical data information, so that they can intuitively understand the traffic and air conditions and their potential links [1].

System environment requirements

Due to high requirement of data warehouse deployment to system hardware environment, the data warehouse of the system is deployed on the X86 server, and for the CISC (Complex Instruction Set Computer) type, the memory should be at least 16GB.

The development platform of the system is Eclipse4.3, and the language is Java and Python. Java language and Python language can be integrated in the Eclipse platform, wherein the Python language can call many classes in Java and has powerful function for text information processing and data analysis, so it is easier to use than Java. The data warehouse construction platform is Red Hat Enterprise Linux Server release6.4. This system version is stable with comprehensive functions, so it is the best choice for deploying data warehouse.

Work mode of topical crawler

According to the division of collaboration, the work mode of topical crawler can be divided into three types:

(1) Independent work: each crawler works independently and does not communicate with each other. Therefore, it will bring very large repeated webpages collection. The usual way is that different crawlers use completely different initial Url lists to minimize the possibility of repeated acquisition.

(2) Dynamic allocation: there is a logical central coordination node to divide internet webpage collection task into smaller parts through a particular partitioning algorithm and dynamically allocate them to each crawler node. Each crawler returns the connection extracted from the webpage to the central coordination node, and then the central coordination node performs the unified dynamic task allocation. This can avoid repeated collection of webpages and dynamically adapt to the network situation. However, with the increase of data acquisition amount, the communication and data interaction between each sub node and central coordination node will become more and more frequent, so the central coordination node become the bottleneck of the whole system [2].

(3) Static allocation: there is also a logical central coordination node, and the acquisition task is also allocated to different crawlers. However, the partitioning method is static and given in advance. Thereby, when each crawler analyzes the underlying link, it does not upload the data to the central coordination node. Thus, the central coordination node only performs the functions of initialization, monitoring or special condition processing. Therefore, with the increase of data, the communication between the sub node and the central coordination node will not significantly increase and avoid the central coordination node becomes the bottleneck of the system.

Among the above three work modes, the static allocation mode can be divided into the following three categories according to different webpage link delivery modes:

a. Firewall mode: in this mode, the local crawler only crawls the webpage indicated by internal links (in the task partitioning way of static allocation mode, the indicated webpage belongs to the local collection task link) without collecting nor transferring the webpage indicated by external links (in the task partitioning way of static allocation mode, the indicated webpage does not belong to the local collection task link). In this way, the communication between different crawler nodes

can be reduced. However, it will lose a large number of extracted links and a lot of webpages can not be successfully collected.

b. Cross mode: in this mode, the local crawler not only collects the webpages indicated by internal links but also collects the webpages indicated by external links. Therefore, the parallel crawler system using this mode will collect more webpages than that using the firewall mode. However, there is no communication interaction among the local crawlers. Therefore, there will be a great deal of repeated acquisition.

c. Interactive mode: in this mode, the local crawler only collects the webpages indicated by internal links but not collects the webpages indicated by external links, and sends the acquisition request of external link to the crawler node of the acquisition task. This mode prevents losing the link of webpage and avoids repeated collection. However, when the system runs, different crawler nodes interact URL acquisition task information with each other. Therefore, it will bring some communication overhead to the system [3].

Technical application of data warehouse

The Online Analytical Processing (OLAP) is one of the main applications of data warehouse technology. The concept of OLAP was first proposed by E.F.Codd, the father of relational databases, in 1993. At that time, Codd thought that Online Transaction Processing (OLTP) could not meet the needs of users for database query analysis, and the simple query of large databases by SQL could not meet the needs of user analysis either. The decision analysis of user requires a large amount of calculation of the relational database to obtain the results, while the query results cannot meet the requirements of the decision maker. Therefore, Codd puts forward the concept of multidimensional database and multidimensional analysis, namely OLAP. The table below is comparison between OLAP and OLTP.

Table 1 Comparison of Data between OLTP and OLAP

OLTP Data	OLAP Data
Raw data	Derived data
Detail data	Comprehensive and extracted data
Current data	Historical data
Updateable	Non-updateable but periodically refreshable
Small data volume processed at a time	Large data volume processed at a time
Application-oriented and event-drived	Analysis-oriented and analysis-drived
Directed to operators and support daily operations	Directed to decision analyst and support management requirements

The OLAP analysis has the following characteristics:

(1) Rapidity: The user has high requirements for fast response capability of OLAP. The system should be able to respond to most of the user's analysis within 6 seconds. If the end user does not get the system response within 30 seconds, it will become impatient and may lose the main line of analysis, which will affect the quality of the analysis. For the analysis of large amounts of data, it is hard to achieve this speed. Therefore, more technical support is needed, such as specialized data storage format, a large number of prior calculations, and special hardware design.

(2) Analyzability: The OLAP system should be able to handle any logic analysis and statistical analysis related to its application. Although the system needs to be programmed in advance, it does not mean that the system defines all of the applications. Users can define new specialized computing without programming, use it as part of the analysis, and give a report in a user's ideal way. Users can perform data analysis on the OLAP platform, and can also connect it to other

external analysis tools, such as time series analysis tools, cost analysis tools, unexpected alarms, data mining, etc.

(3) Multidimensional property: The multidimensional property is the key attribute of OLAP. The system must provide multidimensional analysis and analysis of data analysis, including full support for hierarchical and multi-dimensional dimensions. In fact, multidimensional analysis is one of the most effective methods for analyzing enterprise data, and it is the soul of OLAP[4].

(4) Informativeness: Regardless of the amount of data and the data storage location, the OLAP system should be able to obtain information in time and manage large capacity information. There are many factors that need to be considered, such as data replication, available disk space, performance of OLAP products, and combination with data warehouse.

Analysis on functional requirements of system

The main function modules of the system include: directional data acquisition, data storage and processing, data analysis and display.

(1) Data acquisition: users enter legal web address in the system interactive interface, and enter the specified keywords in the keyword search column to crawl the relevant content of the specified page.

(2) Data preprocessing: users use the ETL tool to select different cleaning rules to process the crawled data, and then the system automatically processes the data and loads it into the data warehouse.

(3) API document parsing: users use the document parsing tool to obtain the given API interface data, and then process and store them into the data warehouse.

(4) Data storage: users use script language to save the extracted data into the data warehouse according to certain rules. The architecture of the module includes log document records and data storage processing.

(5) Data analysis: users analyze the data in the data warehouse through the K mean (K-means) algorithm, the K center algorithm or the FP-growth algorithm button of association analysis.

The user can operate the data in the warehouse through the graphical interface, such as data query, storage and other operations and can also be in the warehouse of the data extraction, using different algorithms for the analysis of data mining, data mining, data analysis, and provide help for decision-making of users [5].

Conclusion

The design purpose of this paper about the environmental pollution is to achieve a unidimensional directional data comparison. By using the system, users can store the data crawled online and the data collected through the API interface into the data warehouse by a preprocessing way, and analyze and present the crawled data to the user through the data mining algorithm. This will be a set of software especially designed for directional data acquisition, intelligent mining and decision analysis for air pollution prevention and control developed by researchers. This system can provide data support for further modification and establishment of effective environmental laws and regulations.

References

[1] Yang Chao, *Research On Large-Scale Web Collection Technology Based On Grid* [dissertation] Harbin, Harbin Institute of Technology, 2007.

[2] Cheng Genshang, Zheng Hongyuan, Ding Qiulin, *A Method of Designing Standard ETL Tool,* Application Research of Computers, Vol.22, No.3, 2005, Page 101-103.

[3] Yan Duanwu, Wang Rifen, *Data Organization and OLAP Implementation of Data Warehouse* Information Science, Vol.21, No.11 2003, Page 1217-1220.

[4] Dong Chao, *Design and Implementation of Vertical Search Engine Based on Topic Information Service* [dissertation], Beijing University of Posts and Telecommunications, 2010.

[5] Zhang Weiming, *Principle and Application of Data Warehouse,* Publishing House of Electronics Industry, 2002.