ATLANTIS PRESS

Advances in Engineering Research, volume 128

7th International Conference on Manufacturing Science and Engineering (ICMSE 2017)

Improvement Research of the Software of Transforming Semi-Structured Html File into Structured Text File

Qiming Cui[1,a], Xue Wang[1,b], Guodong Chen[2,c],Yongbin Zhao[3,d],Bo Li[1,e], Yi Ning[1,f],Shuting Cui [4,g], Zirong Zhang[1,h], Rui Zhao[1,i] ,Hongyu Meng[1,j], Yao Zhang[1,k],Zhenqiang Fu[1,l]

[1]State Grid Anshan Electric Power Supply Company,Liaoning,Anshan 114001,China

[2]State Grid of China Technology College, Shandong,Taian,271000,China

[3]State Grid Liaoning Electric Power Supply Co.,Ltd. Liaoning,Shenyang,110004,China

[4]Eastern Michigan University,Ypsilanti,MI 48197 Michigan,the U.S.A

[a]ospb2002@sina.com,[b]15904920181@163.com,[c]cgdta@126.com,[d]zhaoyb@ln.sgcc.com.cn,[e]61813020@qq.com,[f]272702805@qq.com,[g]dreamilk@126.com,[h]zxcvbnm9172@126.com,[i]lnasfu@163.com,[j]282251716@qq.com,[k]29402147@qq.com,[l]fuzhenqiang1110@163.com

Keywords: Big data; html file; text file;Expert system shell Pro/3;File scan and transformation;Java

Abstract. An application research work had improved some functions of a file scan and transformation software (FileScanner) in Pro/3(an expert system shell) by exploring transformation of semi-structured files (Html format) into structured text files. Some existing problems such as line feed failure and Chinese characters incorrectly displaying in the result file transformed had been solved by improving its Java programming. After a .Html format file be scanned and transformed,a .txt format file produced can implement effectively line feed when it is directly opened, and can display correctly Chinese characters. The structured text file transformed can directly interact with other application programs or databases so as to facilitate the analysis of semi-structured data and mining some values of the information behind the data.

Introduction

At present,many practices relation to the big data become almost the most innovative significance action in technical and business of all industry[1]. when people pay great attention to some values of the big data, some challenges of new are followed with it. In addition to the structured data of the traditional database, there exists a lot of semi-structured data such as Html format in the data. Generally, only after solving the processing difficulties of the semi-structured data, the values of hidden in these data can be effective mined. To some extent, transforming semi-structured data into structured data is a preparation stage of the big data application research. Therefore, exploering this transformation is of great significance.An expert system shell Pro/3 provides a FileScanner program (an executable java-application:net.ligaya.p3.filescanner.jar), which scans and transforms semi-structured (table) data files (for example, in Html-format) into structured data files[2].But the FileScanner has problems in processing line feed and Chinese characters displaying in the file,so some functions of the FileScanner had been improved by researching the FileScanner.

The brief introduction of the FileScanner

The scan and transformation of the FileScanner are based on a format file which specifies the syntax of the files to be scanned as well as the syntax of the generated file[2]. The FileScanner involves some files: a list file to be scanned such as files.tmp, a format file such as html.3ff, a output file transformed such as out.3dl and a fscanner.log of trace-log file,etc.

As a record(one record per file), the name of each file to be scanned such as AuthorList_D.html is stored in the list file in advance. The format file specifies the syntax of the input file and how the scanned components from the file are outputted to the output file. The general format of the format file is described as follows:

tag record

tag record
... ...
-----nodes-----
node record
node record
... ...

Scanning and transforming are carried out according to the following steps(see Fig.1).

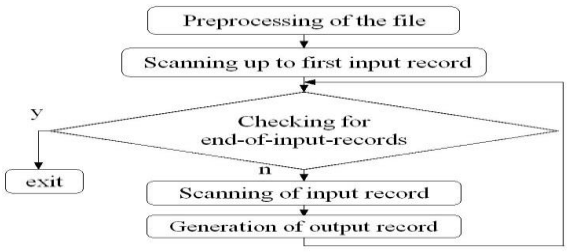

Fig.1 the flow chart of scanning and transforming

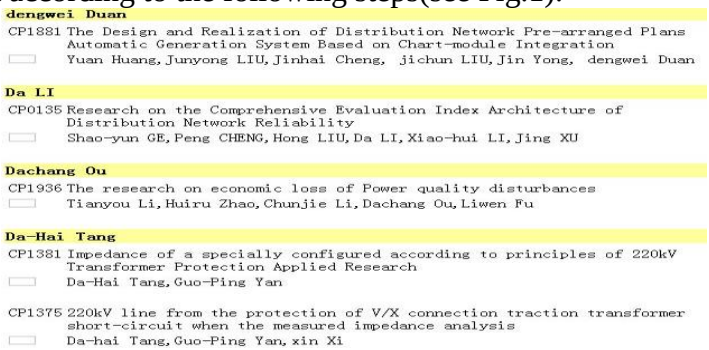

Fig.2 the AuthorList_D.html to be transformed

Tag records have the general format:[tag]=tag_definition tag_definition…. A node record defines a node:<node_name> = format.The preprocessing, input and output record(and so on) include a number of options.

The problems description and analysis of the original FileScanner

The Fig.3 shows the part result of the original FileScanner transforming the Html file shown in the Fig.2. It can be seen that there exist some problems in result file generated by the original FileScanner in windows environment .The result file contains many 0 characters, and implicitly includes a lot of empty(or spaces) lines.if the result file is .txt format file, the line feed doesn't work when it is opened directly by clicking on it. if it is .txt file and it contains some Chinese characters, some unreadable codes will be seen when it is opened directly.if it is opened by other tool such as Word, manually selecting encode mode will be needed.These problems is difficult for reusing the result file converted such as directly(automatic) interacting with other application or database.

The reasons of the above problems are: 1) some 0 characters and empty lines are generated in the processing of transformation; 2) '\n' results in line feed fuction failure; 3) the unreadable codes are caused by the inconsistent of the original encoding (UTF-8) of file to be transformed with the default system file encoding (GBK) (the default system encoding makes the result file transformed becoming GBK codes ,which is inconsistent with the original encoding (UTF-8) of file to be converted).The key to solve the problem of unreadable codes is to ensure the encoding used in conversion consistent with the original encoding of file to be converted. The above problems can be solved by modifying FileScanner and adding the relation function to it.The Fig.4 shows the part result of the modified FileScanner to convert the original Html file shown in the Fig.2.

The improvement of the original FileScanner

The "\n" is replaced by "\r\n" at corresponding to point of line feed failure in the original FileScanner so that the line feed problem will be solved. Similarly, all 0 characters will be removed. The improved script at Transform(String sIn) of FileScan.java[3]in the original FileScanner is as follow:

if(bRecErr){ … … … } Else {sbOut.append(GenRecord(cTagOut.cFirstTagDef, cFormat.cFirstNode, cFormat)).append("\r\n").toString() ;making all new line effective.
sbOut.delete(sbOut.length()-4,sbOut.length()-3) } ; removing all 0 characters.

For the processing of the unreadable codes: First of all, the encoding mode of the original Html file to be converted need to be got. In general, only GBK and UTF-8 encoding mode need to be considered in Chinese windows operating system. The second,for the file of UTF-8 encoding,the values of its three head bytes are -17、-69、-65[4],which can decide whether the encoding of the file is GBK or UTF-8.The problems processing script is indicated as follows(void WriteFile() is added in the FileScan.java)：

```
    void WriteFile()
     { try {RandomAccessFile testFile = new RandomAccessFile("temp file: out.3dl ", "r");
            String testLine = testFile.readLine();
                    while (testLine == null)
           try {String testLine = testFile.readLine(); }
             String testLine1 = testFile.readLine(); int j=0;
             while ( testLine1.charAt(j) == ' '  )
             try { j=j+1; }
              testFile.close();
            OutputStreamWriter Twrite = new OutputStreamWriter(new FileOutputStream("result file:result.txt "));
             boolean bEOF = false; String encode;
             File file = new File("file to be transformed");
             InputStream input= new java.io.FileInputStream(file);
             byte[] byte = new byte[3];
             input.read(byte);
             input.close();
             if (byte[0] == -17 && byte[1] == -69 && byte[2] == -65)
                    encode = "UTF-8";
                    else
                encode ="GBK";
BufferedReader ffFile = new BufferedReader(new InputStreamReader(new FileInputStream("temp file:out.3dl "), encode))              ;encode is the encoding mode of the file to be transformed
       try { while(!bEOF)
                  try {String sLine = ffFile.readLine();
                            if(sLine != null  ){ Twrite.write(sLine.trim());
                    if(sLine.isEmpty()| sLine.length()<=j) continue;
                       Twrite.write("\r\n");}
                     else { bEOF = true;}}
                  try {ffFile.close();Twrite.close(); }
           } } }
```

The example of executing transformation

With reference to the literature[2], the content of format file (Html.3ff) used in the application research is as follows:

```
[SUB]="</td>" ";"
[SUB]=" " ""
[SKB]="<" ">"
[NUL]="" ";"
[SIR]=";"
[EIR]=WL ";" EOF
[INP]=WL <one> L <two>
[OHD]="" L
[OUT]=<one> L <two>  L
----- nodes -----
<one>=aN
<two>=i.
```

The SUB,SKB,NUL,SIR,EIR,INP,OHD and OUT are tag name in the format file,The record of begin with [SUB],[SKB],[NUL],[SIR],[EIR],[INP],[OHD] and [OUT] are tag record. one in the <one> is node name,and the <two> is the same as the <one>.The record of begin with <one> and <two> are node record.

Executing the FileScanner from the command line（or invoking it from the Pro/3）is as follow:

Java -jar D:\ pro3\filescanner.jar D:\ pro3\files.tmp 3ff D:\ pro3\fscanner.log D:\pro3\fscanner.err /OT=D:\ pro3\out.3dl /FF=D:\ pro3\

Where Java -jar D:\ pro3\filescanner.jar is executable jar,namely FileScanner; files.tmp is list file(content: AuthorList_D.html);3ff is html.3ff， namely format file; out.3dl is intermediate file transformed;/FF pointing to the format file 's directory.

```
dengwei Duan 0  0 CP1881 0  0 The Design and Realization of Distribution Network Pre-arranged
Plans Automatic Generation System Based on Chart-module Integration 0  0  0 Yuan Huang,Junyong
LIU,Jinhai Cheng, jichun LIU,Jin Yong, dengwei Duan 0  0  0  0  0 Da LI 0  0 CP0135 0  0
Research on the Comprehensive Evaluation Index Architecture of Distribution Network
Reliability 0  0  0 Shao-yun GE,Peng CHENG,Hong LIU,Da LI,Xiao-hui LI,Jing XU 0  0  0  0  0
Dachang Ou 0  0 CP1936 0  0 The research on economic loss of Power quality disturbances 0  0
0 Tianyou Li,Huiru Zhao,Chunjie Li,Dachang Ou,Liwen Fu 0  0  0  0  0 Da-Hai Tang 0  0 CP1381
0  0 Impedance of a specially configured according to principles of 220kV Transformer
Protection Applied Research 0  0  0 Da-Hai Tang,Guo-Ping Yan 0  0  0  0 CP1375 0  0 220kV line
from the protection of V/X connection traction transformer  short-circuit when the measured
impedance analysis 0  0  0 Da-hai Tang,Guo-Ping Yan,xin Xi |
```

Fig.3 the part transforming result file(out.3dl)of the original FileScanner

```
dengwei Duan
CP1881
The Design and Realization of Distribution Network Pre-arranged Plans Automatic Generation
System Based on Chart-module Integration
Yuan Huang,Junyong LIU,Jinhai Cheng, jichun LIU,Jin Yong, dengwei Duan
Da LI
CP0135
Research on the Comprehensive Evaluation Index Architecture of Distribution Network
Reliability
Shao-yun GE,Peng CHENG,Hong LIU,Da LI,Xiao-hui LI,Jing XU
Dachang Ou
CP1936
The research on economic loss of Power quality disturbances
Tianyou Li,Huiru Zhao,Chunjie Li,Dachang Ou,Liwen Fu
Da-Hai Tang
CP1381
Impedance of a specially configured according to principles of 220kV Transformer Protection
Applied Research
Da-Hai Tang,Guo-Ping Yan
CP1375
220kV line from the protection of V/X connection traction transformer  short-circuit when the
measured impedance analysis
Da-hai Tang,Guo-Ping Yan,xin Xi
```

Fig.4 the part transforming result file(result.txt)of improved FileScanner

Discussion

The list file is an ASCII file with one record per file to be scanned,namly,one or multiple files can be scanned each time.The content of the format file is very important.If some settings of the format file is inappropriate, the transformation result is very difference. The node name is a character string. Node naming is no special restrictions,but it must be consistent and not be omitted in the format file. The line feed doesn't work in the result file transformed by the original FileScanner ;if there are any Chinese characters in the file to be transformed ,some unreadable codes will be generated, which is not convenient to directly doing subsequent processing in the expert system shell pro/3 or other application(although the FileScanner embeds in Pro/3,but it does not rely on Pro/3, it can be run separately) ,and more manual processing need to be done.Through the improvement, the result.txt solving these problems is generated on the basis of the original output file out.3dl, which can be easily doing subsequent processing in the expert system shell pro/3.

Conclusions

The paper work had researched the transformation of semi-structured data file (such as .html format) into structured data file (such as .txt format) by improving some functions of the original FileScanner in the expert system shell Pro/3.The result of the research include:1)Removing all 0 characters, empty(or spaces) lines in result file generated by the original FileScanner during scanning and transforming;2)Making the .txt format file transformed correctly line feed when it is opened directly by clicking on it,which is basic of automatic interact with application ;3)According to the encoding mode of the original files to be transformed, automatically adjusting the encoding mode of the .txt file to be generated such that the .txt format file containing Chinese characters does not generate unreadable codes when it is directly opened by clicking on it;4)The result file transformed is structured such that doing subsequent processing is easily in the expert system shell pro/3,and directly interacting with other application or database is convenient and can be easy automatic.

References

[1]Shishan Gu. unstructured data analysis: a new value of big data era, http://soft.chinabyte.com/308/12848808.shtml

[2]Jens Hintze Holm, Pro/3 expert system shell main page, http://www.ligaya.net/p3/FileScanner.html

[3]Jens Hintze Holm, net.ligaya.p3.filescanner.jar
Available:http://www.ligaya.net/p3/download.html
[4]SayGoodbyeToYou,http://blog.csdn.net/saygoodbyetoyou/article/details/11921909