AUTOMATIZED WEB PAGES PARSING AND CREATION

Boris Bocharov, Maria Voevodina

O. M. Beketov National University of Urban Economy in Kharkiv

No subject area can be automated by one hundred percent. Some operations inevitably continue to be performed in a "manual" version, for others, as far as changes occur with the object of automation, you still have to make additional programs. It happens that even a very simple task can require writing hundreds of lines of code. In addition, absolutely standard and one-type operations with files require a lot of time-consuming and painstaking work.

Unfortunately, the remarkable ideas implemented in the UNIX operating system are forgotten -– the creation of small programs "on the fly," the data pipelining, the use of specialized programs for working with files.

In this article, we tried to apply these concepts to improve the efficiency of work (creating, editing, analyzing) with web pages using the AWK program.

The AWK program was created almost 40 years ago and was included in all versions of UNIX. The name of the program is the first letters of the authors' names (Aho, Weinberg, Kernigan). We will use the GNU (Gnu's not UNIX) version of the program, which is distributed free of charge. The program has been moved to the Windows operating environment. It's called gawk.exe.

The article is aimed at programmers who know the C language or one of its clones (JavaScript, PHP, etc.).

O. M. Beketov National University of Urban Economy in Kharkiv pays great attention to Internet technologies in various fields of activity. The site of the Applied Mathematics and Information Technologies Department has been created and is being actively used.

It is very important to have information about the effectiveness of each page. Unfortunately, we do not have the right to access system information (the department site is part of the university's website, which is created on the Joomla platform). In these conditions, the simplest and most reliable way to determine the number of visits to each page of the site is to analyze the materials manager service information.

We used the AWK program [1-4] for parsing the pages.

The following information is extracted:

- page name,
- page category,
- page number of visits,
- page creation date,

- page id.

Records and fields jhnions:
"</tr>" – records separator,
"</td>" – fields separator.

Service information for parsing:
title="Change"> – the page name string prefix,
**** – the page name string suffix,
Category: – the page category string prefix,
** ** – the page category string suffix,
**** – the page number of visit string prefix,
**** – the page number of visit string suffix,
<td class="nowrap small hidden-phone"> – the page creation date string prefix,
<td class="hidden-phone"> – the page id string prefix.

The program text is presented below.

```
# read information from file and write in the array (key => value)
function read_strings(fn,srt_arr,    s_key,s_val,r){
  while(r = getline s_key < fn){
    if(r == -1){
      print "ERROR: " fn " - " ERRNO;
      return -1;
    }
    r = getline s_val < fn
    srt_arr[s_key] = s_val;
  }
  return 0;
}

# removing extra spaces
function trim(str){
  trim_str = str;
  gsub(/^ +/,"",trim_str);
  gsub(/ +$/,"",trim_str);
  gsub(/ {2,}/," ",trim_str);
  return trim_str;
}
```

```
# replacement of service characters with spaces
function srv2space(src,    s){
  s = src;
  gsub(/\n/," ",s);
  gsub(/\t/," ",s);
  return s;
}

# deletion of a prefix
function cut_left(src,fnd,   n,s){
  n = index(src, fnd);
  if(n==0) return "";
  s = substr(src,n+length(fnd));
  return s;
}

# deletion of a suffix
function cut_right(src,fnd,   n,s){
  n = index(src, fnd);
  if(n==0) return "";
  s = substr(src,1,n-1);
  return s;
}

# deletion of prefix and suffix
function cut_both(src,le,ri,  n,s){
  if(le!="") s = cut_left(src,le);
  if(s==0) return "";
  if(ri!="") s = cut_right(s,ri);
  return s;
}

# BEGIN section of program
BEGIN{
  read_strings("info.txt",G);
  RS="</tr>";
  FS="</td>";
}

# MAIN section of program
{
  nm="";
```

```
  kt="";
  ps="";
  dt="";
  id="";
  for(i=1; i<=NF; i++){
# name
    s=cut_both($i,G["N1"],G["N2"]);
    if(s != ""){
      s = srv2space(s);
      nm = trim(s);
    }
# category
    s=cut_both($i,G["K1"],G["K2"]);
    if(s != ""){
      s = srv2space(s);
      kt = trim(s);
    }
# number of visits
    s=cut_both($i,G["P1"],G["P2"]);
    if(s != ""){
      s = srv2space(s);
      ps = trim(s);
    }
# creation date
    s=cut_both($i,G["D"],"");
    if(s != ""){
      s = srv2space(s);
      dt = trim(s);
    }
# id
    s=cut_both($i,G["ID"],"");
    if(s != ""){
      s = srv2space(s);
      id = trim(s);
    }
  }
  if((nm!="")||(kt!=""))
    printf("%s\t%s\t%s\t%s\t%s\n",nm,kt,ps,dt,id);
}
```

The following bat-file calls the parsing program

```
gawk --re-interval -f mk.awk src.html >res_utf.txt
```

src.html – html-file saved to a local disk,
res_utf.txt – result (text file in utf-8 encoding).

Result file fragment is presented below.

Dep. of PM and IT	About the department	21705	07.09.15	18
Department Staff	About the department	4676	26.11.13	16
Conferences	Science	3332	17.12.13	13
Scientific directions	Science	3143	17.12.13	12
Consultations	For students	561	17.12.13	11
Distance learning	For students	757	16.12.13	10

References:

1. Бочаров Б.П. Інформаційні технології в освіті : монографія / Б.П. Бочаров, М.Ю. Воєводіна; Харків. нац. ун-т міськ. госп-ва ім. О. М. Бекетова. – Харків: ХНУМГ ім. О. М. Бекетова, 2015. – 197 с.
2. Бочаров Б.П. AWK - универсальная программа работы с текстовыми файлами / Б.П. Бочаров, М. Ю. Воеводина // Библиотеки учебных заведений.– 2002.– №4.– С. 39-53.
3. Бочаров Б.П. Глобальная корректировка БД с использованием программы AWK / Б.П. Бочаров, М.Ю. Воеводина, Л.П. Семененко // Культура народов Причерноморья.– 2003.– № 40.– С. 78-81.
4. Бочаров Б.П. Формирование отчетов в электронных каталогах / Б.П. Бочаров, М.Ю. Воеводина // Библиотеки учебных заведений.– 2003.– № 10.– С. 41-61.