AUTOMATIC E-MAIL DOWNLOAD THROUGH WWW INTERFACE

Matej Kuna
Bachelor Degree Programme (3), FIT BUT
E-mail: xkunam01@stud.fit.vutbr.cz

Supervised by: Dušan Kolář
E-mail: kolar@fit.vutbr.cz

ABSTRACT

This paper deals with automatic access a service that is accessible only through a www interface. I created a set of programs in Python programming language. These programs are able to download e-mails from the operamail.com server, without user assistance. I am working on other servers now.

1. ÚVOD

Prístup k elektronickej pošte alebo vo všeobecnosti k službám, ktoré sú prístupné iba cez www rozhranie môže byť pre užívateľa z viacerých dôvodov nevhodné. Jedná sa o prípady, keď chce užívateľ získať textové informácie a presne vie, kde má tieto informácie hľadať.

Výborná ukážka je napríklad prečítanie všetkých emailov z emailového konta, ktoré je prístupné cez www rozhranie. Ide tu o činnosť, ktorú je možné vo viacerých prípadoch automatizovať. Z toho dôvodu som sa rozhodol vyvinúť sadu programov v skriptovacom jazyku Python, ktorá umožňuje automatické stiahnutie mailov zo serveru operamail.com.

Vo všeobecnosti sa orientujem na rôzne bezplatné emailové servery. Výsledky mojej analýzy ukázali, že spravidla ponúkajú bezplatný prístup k emailovej schránke len cez www rozhranie. Prístup cez protokoly POP3 a IMAP je tiež možný, ale je spoplatnený.

2. POUŽITIE

Program je možné predovšetkým využiť v prípadoch keď chceme pristupovať k službe cez www rozhranie a sú splnené podmienky.

- k službe sa treba prihlásiť cez HTML formulár
- a (alebo) je potrebné počas doby prístupu udržiavať stav prostredníctvom HTTP protokolu (cookies)
- zo súborov prenesených cez HTTP protokol potrebujeme iba určitú časť textu, ktorá sa uloží na disk

Program by samozrejme dokázal uložiť na disk aj celý HTML dokument, nie len jeho časť, ale na takéto prípady existujú iné nástroje.

3. ČINNOSŤ PROGRAMU

V jednoduchosti vysvetlím ako program pracuje a čo všetko je potrebné vykonať pred samotným spustením.

3.1. NASTAVENIA

Aby program mohol automaticky vykonávať činnosť, ktorú užívateľ potrebuje, je nutné najprv previesť nastavenia. Nastavenia sú uložené v konfiguračnom súbore a o ich správu sa stará samostatný program.

Nastavenia je možné rozdeliť do dvoch skupín. Prvou z nich je skupina základných nastavení, kde patrí adresa serveru a prihlasovacie údaje (ak sú potrebné).

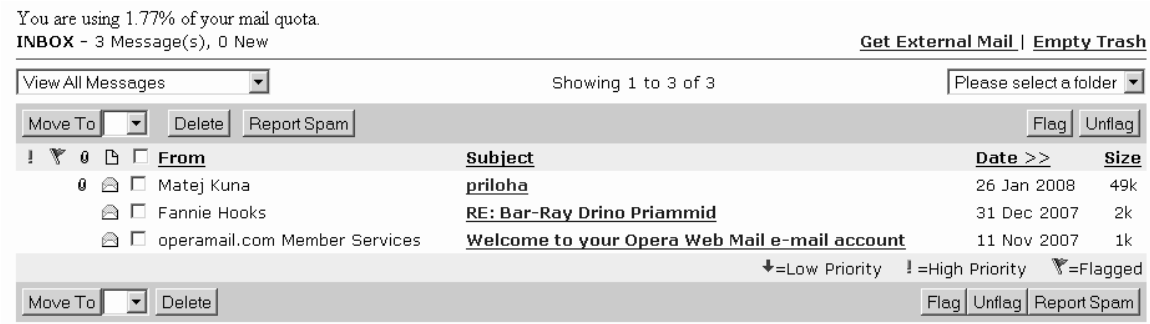
Druhú skupinu tvorí súbor pravidiel. Tie by sa dali prirovnať k veľmi primitívnemu skriptu. Ich syntax sa nepodobá žiadnemu známemu programovaciemu alebo skriptovaciemu jazyku. Určujú, čo má program robiť po pripojení sa k HTTP serveru. Každé pravidlo sa začína číslom, aby bolo možné medzi mini skákať v prípade potreby. Pravidiel je približne 10 a ich syntax nebudem ďalej rozoberať.

3.2. POŽADOVANÁ AUTOMATICKÁ ČINNOSŤ

Ak sú už všetky potrebné nastavenia uložené, na rad príde ďalší samostatný program, ktorého úlohou je interpretácia pravidiel. V tejto fáze sa komunikuje s HTTP serverom a na základe pravidiel sa mu zasielajú požiadavky. Program pri tejto činnosti simuluje internetový prehliadač užívateľa. Znamená to, že otvára hypertextové odkazy a ukladá si do pamäte získaný HTML dokument (prípadne iný formát). Počas behu samozrejme udržiava stav prostredníctvom cookies, dokáže korektne načítať HTML stránku, ktorá obsahuje rámy (frames) alebo automaticky načíta inú adresu pri presmerovaní.

3.3. VÝSLEDOK

Obrázok 1 ukazuje webové rozhranie serveru operamail.com. Je vidieť že v schránke sú 3 správy a z toho prvá správa obsahuje prílohy. Na obrázku 2 sú pravidlá, ktoré uložia všetky e-maily vrátane príloh.



Obrázok 1: www rozhranie prijatých správ na serveri operamail.com

```
1 login
2 link 'go to inbox'i -> inbox
# otvorenie prvého e-mailu
3 inbox->page r'.*mail\/read\.mail\?folder\=INBOX\.*'i-> tmp |6,80
# vybratie hlavičiek a tela e-mailu z HTML dokumentu
6 grep 'from\:\s*(.*?)\s*\['i -> from
7 grep 'subject\:\s*(.*?)\<'i -> subject
8 grep 'D ml_header\.htm \-\-\>(.*?)\<\!\-\- START ml_foot'-> mail
# cyklus na ukladanie príloh (začiatok)
15 mail -> grep '\"BLUE\"\>(.*?)\,?\s*\&' -> attach_name |20,25
20 mail -> page r'.*getatt.*' -> attachment
21 page r'.*getattach.*' -> attachment | 22,22
22 write attachment -> subject + '_' + attach_name
23 goto 15
# cyklus na ukladanie príloh (koniec)
24 write 'From: ' + from + '\n' + 'Subject: ' + subject -> subject
+'.txt'
25 write '\n\n' -> subject +'.txt'
26 write mail -> subject +'.txt'
27 close -> subject +'.txt'
# prechod na ďalší e-mail
28 tmp -> link 'next'i -> tmp |6,80
80 stop
```

Obrázok 2: pravidlá programu pre server operamail.com

4. ZÁVER

Vytvorený program dokáže automaticky uložiť všetky emailové správy zo serveru operamail.com. Tieto nastavenia zahŕňajú aj uloženie príloh, ktoré môže mať každá správa. Predpokladám, že modifikáciou pravidiel bude možné ukladať emaily z ľubovolného serveru, ktorý ma podobné užívateľské rozhranie ako operamail.com.

Táto sada programov bola primárne tvorená a testovaná pre automatické sťahovanie emailov. Vytvorením nových pravidiel bude možné automatizovať aj iné potrebné činnosti. Môže sa jednať napríklad o ukladanie príspevkov z internetovej diskusie na disk. Tieto príspevky môže následne iný jednoduchý program zasielať užívateľovi na email, ku ktorému bude pristupovať cez emailového klienta.

V blízkej dobe sa zameriam na napísanie pravidiel, ktoré by umožňovali automatické stiahnutie emailov zo známych mailových serverov domény .cz.

REFERENCES/LITERATURA

[1] Python Documentation [online]
URL: <http://www.python.org/doc/> [cit. 2008-03-03]

[2] RFC 2109 - HTTP State Management Mechanism [online]
URL: <http://www.faqs.org/rfcs/rfc2109.html> [cit. 2008-03-03]

[3] RFC 2616 - Hypertext Transfer Protocol – HTTP/1.1 [online]
URL: <http://www.faqs.org/rfcs/rfc2616.html> [cit. 2008-03-03]