VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Síťové aplikace a správa sítí – projekt

Čtečka novinek ve formátu Atom a RSS s podporou TLS

11. novembra 2022

Juraj Dedič

Obsah

1 Novinové zdroje

Úlohou bolo vytvoriť program na sťahovanie a zobrazovanie novín vo formáte RSS 2.0 alebo Atom.

2 Implementácia

Zvolený implmentačný jazyk je C++, s pomocou knižníc openssl a libxml2.

2.1 Spracovanie argumentov

Spracovanie argumentov je riešené pomocou funkcie getopt. Pomocou tejto funkcie program zachytí prepínače pre zobrazenie autora, času a URL adresy článku.

Pre zobrazenie zvoleného zdroja je možné zadať URL zdroja ako argument programu.

Program akceptuje možnosť zadania umiestení zdrojov do tzv. `feedfile`. Umiestenie feedfile je špecifikované prepínačom `-f`.

Na poradí argumentov nezáleží, s vínimkou argumentu pre -f, za ktorým musí nasledovať umiestnenie súboru feedfile.

2.2 Príprava

Po spracovaní argumentov program pripraví dočasný adresár `./temp/` pre sťahovanie zdrojov.

Následne program spustí sťahovanie a výpis nad URL zadanou v argumentoch. V prípade využitia feedfile, je využitý cyklus ktorý prechádza tento súbor po riadkoch. Pokiaľ je riadok neprázdny a zároveň nieje daný riadok komentár, program spustí sťahovanie a výpis nad s URL adresou na tomto riadku. Cyklus takto ďalej pokračuje až pokým sa nedostane na koniec feedfile.

2.3 Spracovanie URL

Pred sťahovaním súboru s novinkami je potrebné získať informácie z URL adresy. Toto je riešené pomocou funkcie `parse_url` ktorá vracia štruktúru obsahujúcu hostname, port, umiestnenie zdroja a informáciu, či zdroj využíva HTTPS.

2.4 Sťahovanie

Funkcia `read_from_url`, ktorá je spustená z hlavnej časti programu zavolá príslušnú funkciu z `downloader.hpp`, podľa toho či sa jedná o zdroj využívajúci HTTP alebo HTTPS.

V prípade, že sa jedná o HTTPS, program zavolá inicializačnú funkciu, ktorá je vyžadovaná pre kompatibilitu s verziou openssl nižšou ako 1.1. Program následne pomocou makra z dokumentácie openssl zvolí TLS metódu, kde pri starších verziách knižnice bude použitá metóda TLS 1.2. Nový SSL_CTX objekt umožňujúci TLS spojenie bude potom využívať túto metódu.

Pokiaľ užívateľ špecifikoval cestu k adresáru alebo súboru s certifikátmi, bude použitá funkcia pre načítanie týchto certifikátov. Pri absencii týchto parametrov, bude využité úložisko certifikátov poskytnuté funkciou `SSL_CTX_set_default_verify_paths`. Na základe týchto certifikátov budú verifikované certifikáty poskytnuté serverom.

Ďalej sú nastavené parametre TLS spojenia a šifrovacie algoritmy. Program pokračuje nadviazaním spojenia a dokončením handshaku. Po tomto je získaný a verifikovaný certifikát serveru.

U HTTP bez TLS je postup pripojenia zjednodušený vynechaním funkcií TLS. Pri HTTP je iniciovaný Basic Input/Output (BIO) pre komunikáciu medzi klientom a serverom pomocou funkcie `BIO_new_connect`. Volanie `BIO_do_connect` zaistí zahájenie spojenia.

Po zahájení spojenia aplikácia pošle pomocou HTTP metódy GET požiadavku pre získanie zdroja v zadanom umiestnení a špecifikuje využitie HTTP/1.0[1] a ďalšie potrebné prvky.

Po prijatí odpovede je telo (obsahúce XML) oddelené od HTTP hlavičky. Následne je tento obsah zapísaný do súboru `./temp/temp.xml` a uvoľnia sa alokované prostriedky.

2.5 Parsovanie a výpis

Po stiahnutí súboru je získaný koreňový element. Ak je názov koreňového elementu „rss“, vieme že používa tento formát a preto nastavíme za nový root element s názvom „channel“. V prípade že názov je „feed“, budeme s ním pracovať ako s formátom Atom. Ak ani jeden z týchto názvov nezodpovedá, bude tento súbor preskočený .

TODO

Následne program iteruje cez elementy kanálu a pre každý element „entry“ je volaná funkcia `parse_item`. Táto funkcia postupuje taktiež iteratívne a hľadá nasledujúce elementy:

- title - obsahuje v tele nadpis článku
- link - v prípade RSS 2.0 obsahuje URL článku v parametre href. U formátu Atom je táto informácia uložená priamo v tele prvku
- published - obsahuje dátum publikovania
- author - obsahuje informácie o autorovi (meno a email) v dielčích elementoch.

[1] Táto verzia bola zvolená z dôvodu, že pri použití verzie 1.1 obsahovalo telo odpovede nechcené artefakty. Toto je zrejme spôsobené využitím chunked transfer encoding u verzie 1.1

Literatura