VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií

Čtečka novinek ve formátu Atom s podporou TLS

DOKUMENTACE

Autor: Martin Omacht

Login: xomach00

1 OBSAH

2 ÚVOD

Tento dokument slouží jako dokumentace k programu *feedreader*. Tento program funguje jako čtečka novinek ve formátu Atom nebo RSS s podporou TLS. Dokumentace zároveň poskytuje základní informace o formátu Atom a RSS a také o protokolu TLS/SSL.

3 ČTEČKA NOVINEK

Než se pustíme do implementačních detailů, je potřeba si říct něco o tom, co je to čtečka novinek.

Čtečka novinek je program, který stahuje aktuality z webových zdrojů (angl. *feed*). Tyto zdroje se většinou vyskytují na stránkách s často měnícím se obsahem (např. zpravodajské servery) (1). Nejčastější formáty zdroje jsou Atom a RSS, které využívají XML formát.

3.1 RSS

Zkratka RSS má několik výkladů (1):

- Rich Site Summary (RSS 0.91 od firmy Netscape)
- Resource Description Framework Site Summary (RSS 1.0 – tvůrcem je W3C[1])
- Really Simple Syndication (RSS 2.0 – spravováno Berkman Klein Center for Internet & Society[2])

Formát verzí, které tato čtečka podporuje, a důležité elementy formátu, si rozebereme v následujících sekcích.

3.1.1 RSS 0.91

Kořenovým elementem musí být `<rss>` s povinným atributem `version`, který specifikuje verzi používaného RSS (v tomto případě 0.91). V něm se nachází jediný `<channel>` element, který obsahuje informace o kanálu a jeho obsahu (2). Tento element pak může obsahovat řadu povinných a nepovinných elementů. Pro účely této čtečky je důležitý povinný element `<title>`, který udává název kanálu.

Jednotlivé položky zdroje jsou uvedeny v elementech `<item>`, které se vyskytují taktéž v elementu `<channel>`. Těchto položek může zdroj obsahovat jakékoliv množství. V elementu `<item>` jsou pak povinné elementy `<title>` a `<link>`. Tyto elementy obsahují titulek a odkaz na položku. Informace o autorovi nebo poslední aktualizace položky tento standard nepodporuje.

Čtečka podporuje i verzi 0.92, která pouze přidává několik volitelných elementů pro element `<item>`.

3.1.2 RSS 2.0

Tato verze je založena na verzi 0.91. Jelikož je verze 2.0 zpětně kompatibilní s verzí 0.91, základní struktura elementů je stejná. Pro tuto čtečku však verze 2.0 přidává důležité elementy pro element `<item>`:

[1] https://www.w3.org/
[2] https://en.wikipedia.org/wiki/Berkman_Klein_Center_for_Internet_%26_Society

- <author> - email autora položky
- <guid> - řetězec, který unikátně identifikuje položku. Pokud neobsahuje atribut isPermaLink="false", čtečka obsah tohoto elementu použije jako URL odkaz na položku.
- <pubDate> - datum a čas publikace položky

Standard také změnil všechny dílčí elementy prvku <item> na volitelné, za podmínky, že bude přítomný alespoň <title> nebo <description>. (3)

3.1.3 RSS 1.0

Verze 1.0 je založená na RSS 0.9 a zachovává zpětnou kompatibilitu, ale není kompatibilní s verzemi 0.91 a 2.0, které nepoužívají RDF (Resource Description Framework).

Kořenovým prvkem je element <rdf:RDF>, kde prefix jmenného prostoru rdf je asociován schématu syntaxe RDF (může být použit i jiný prefix, ale pro zpětnou kompatibilitu je doporučené používat rdf). Stejně jako ve výše popsaných verzích RSS tento element obsahuje prvek <channel> popisující samotný kanál. Je zde například povinný element <title>, který obsahuje titulek kanálu.

Na rozdíl od RSS verzí 0.91 a 2.0, elementy <item> nejsou umístěné v elementu <channel>, ale jsou přímo v kořenovém prvku. Elementy <title> a <link> jsou stejné jako v ostatních případech. Elementy pro autora a poslední aktualizaci položky standard nepodporuje, ale jeho oficiální modul Dublin Core[3] povoluje elementy <dc:creator> a <dc:date> pro tento účel (prefix jmenného prostoru dc je asociován modulu Dublin Core). (4)

3.2 ATOM

Atom je další standard pro publikování syndikovaného webového obsahu, který se snaží vyhnout limitacím a chybám standardu RSS (5).

Kořenový element je <feed>. Ten obsahuje mimo jiné název zdroje v povinném elementu <title>. Dále může obsahovat jakýkoliv počet elementů <entry>, které reprezentují jednotlivé položky zdroje.

Titulek položky zdroje se opět nachází v povinném elementu <title>. Poslední aktualizace reprezentuje element <updated>. Co je ale složitější, je autor položky. Ten se vyskytuje v elementu <author>, ve kterém se pak nachází povinně element <name> se jménem autora. Elementů <author> může být uvedeno více nebo žádný a v takovém případě se použije <author> z elementu <source>. Pokud není k dispozici autor ani v <source>, jako autor se považuje autor uvedený v elementu <feed>. Odkaz na webovou stránku položky je obsahem atributu href v elementu <link>, který buď nemá atribut rel nebo atribut rel má hodnotu alternate. (6)

[3] http://web.resource.org/rss/1.0/modules/dc/

4 TLS

TLS (Transport Layer Security) protokol poskytuje zabezpečené připojení mezi dvěma komunikujícími koncovými uzly. TLS vychází z dřívějších protokolů SSL, ve kterých však byly odhaleny bezpečnostní díry, a tak je dnes tento protokol již zastaralý.

Zabezpečené připojení přes TLS poskytuje následující vlastnosti:

- Autentizace – serverová část připojení je vždy autentizovaná; klientská část je volitelně autentizovaná.
- Důvěrnost – odeslaná data jsou vždy viditelná pouze koncovým uzlům.
- Integrita – odeslaná data nemohou být bez detekce modifikována útočníky.

Tyto vlastnosti jsou dodrženy i pokud útočník má k dispozici kompletní kontrolu nad sítí.

TLS se skládá z dvou primárních komponent:

- Handshake protokol – autentizuje komunikující strany, dohodne kryptografické módy a parametry a zavede klíčovací materiál.
- Record protokol – použije dohodnuté parametry pro zabezpečení komunikace mezi koncovými uzly.

Během fáze handshake server odešle certifikát klientovi, který ho ověří proti souboru důvěryhodných certifikátů (autentizace serveru). Klient odesílá certifikát pouze pokud si ho server vyžádá (autentizace klienta). (7)

5 PROGRAM FEEDREADER

Program feedreader je čtečka novinek ve formátu Atom a RSS s podporou TLS. Uživatelským rozhraním je příkazový řádek. Program stáhne na základě parametrů uvedené zdroje, které mohou být ve formátu Atom nebo RSS, a vypíše uživatelem požadované informace na standardní výstup.

5.1 ZÁKLADNÍ INFORMACE

Výstup se skládá z názvu zdroje (kanálu) a jednotlivých položek (novinek). Parametry lze přidat vypisování autora položky, odkazu na položku nebo čas poslední úpravy položky (viz 5.2)Použití. Pomocí souboru předaným parametrem `-f` lze specifikovat více zdrojů.

5.1.1 Podporované formáty zdrojů

Program je testovaných na následujících verzích formátu zdrojů:

- RSS 0.91
- RSS 2.0
- RSS 1.0
- Atom

Avšak díky kompatibilitě některých formátu a benevolenci programu, by neměl být problémy s těmito verzemi:

- RSS 0.9
- RSS 0.92
- RSS 2.0.1-2.0.11

5.1.2 Podporované komunikační protokoly

Program podporuje protokol HTTP i HTTPS pro stahování zdrojů.

5.1.3 Podporované platformy

Program je testován pouze na platformě Linux.

5.2 Použití

Možnosti spuštění:

```
feedreader <URL | -f <feedfile>> [-c <certfile>] [-C <certdir>] [-T] [-a] [-u]
```

Popis parametrů:

Pořadí parametrů je libovolné. Povinně musí být uveden parametr URL, nebo parametr -f. Každý parametr lze zadat pouze jednou.

- URL – URL adresa zdroje novinek
- -f <feedfile> - Určí cestu k souboru s URL adresami zdrojů, formát viz 5.2.1
- -c <certfile> - Určí soubor s důvěryhodnými certifikáty, které budou sloužit k ověření certifikátu serveru
- -C <certdir> - Určí adresář s důvěryhodnými certifikáty, které budou sloužit k ověření certifikátu server. Před použitím adresáře je potřeba adresář připravit pomocí příkazu c_rehash.
- -T – Přidá do výpisu položky zdroje informaci o času poslední úpravy položky.
- -a – Přidá do výpisu položky zdroje informaci o autorovi položky.
- -u – Přidá do výpisu položky zdroje URL odkaz položky.

Při zadání nesprávných parametrů se vypíše nápověda.

5.2.1 Formát souboru feedfile

Soubor feedfile je textový soubor, který obsahuje URL adresy zdrojů, které má program stáhnout. Tyto URL adresy jsou oddělené novým řádkem. Bílé znaky a prázdné řádky jsou ignorovány. Soubor musí obsahovat alespoň jednu URL adresu. Název souboru může být libovolný.

Také je možné použít komentáře. Ty začínají znakem # (křížek). Veškerý text uvedený za tímto znakem až do konce řádku je ignorován. Komentář musí začínat hned na začátku řádku nebo se bezprostředně před ním musí nacházet znak mezery.

Příklad obsahu souboru feedfile:

```
# Základní URL
url.com
                # Komentář s křížkem #

http://www.nic.cz:8484/some/url#id #Komentář
```

5.2.2 Formát výstupu

Název zdroje je uvozený znaky „*** “ a ukončený znaky „ ***“, např. „*** Titulek ***“. Na dalších řádcích jsou pak titulky jednotlivých položek zdroje. Pokud položka nebo zdroj nemá titulek, bude místo něj vypsáno „<< BEZ NÁZVU >>“.

V případě použití parametrů -a, -u nebo -T budou tyto informace vypsané pod titulkem položky a jednotlivé položky budou odděleny prázdným řádkem.

Autor položky je uvozen řetězcem „Autor: “. Pokud autor položky není zdrojem uveden, nevypíše se za tento řetězec nic dalšího. Pokud bude autorů více budou jejich jména oddělené středníkem (například „Autor: Petr Novák; Jana Novotná“).

URL položky je uvozeno řetězcem „URL: “. Pokud URL položky není uvedeno, nevypíše se za tento řetězec nic dalšího.

Čas poslední aktualizace položky je uvozen řetězcem „Aktualizace: “. Pokud poslední aktualizace položky není uvedena, nevypíše se za tento řetězec nic dalšího.

Jestliže bude použit parametr -f, tak jednotlivé zdroje od sebe budou odděleny prázdným řádkem.

5.2.3 Návratové kódy

Program má několik návratových kódu, které indukují určité chyby.

5.2.3.1 Chyba argumentů (1)

Chybový kód 1 značí špatně zadané argumenty aplikace. Tuto chybu může vyvolat například zadání parametru, který není uveden mezi podporovanými parametry, vynecháním povinného parametru nebo duplicitním výskytem parametrů.

5.2.3.2 Chyba připojení (2)

Chybový kód 2 značí chybu při připojení. Toto může znamenat například chybu ověření serverového certifikátu, nemožnost připojení k serveru nebo připojení k internetu není k dispozici.

5.2.3.3 Nepodporovaný HTTP status (3)

Tato chyba nastane, pokud server vrátí HTTP status, který tato aplikace nepodporuje (například 3xx).

5.2.3.4 Chyba HTTP (4)

Návratový kód 4 znamená chybu v přijaté HTTP odpovědi.

5.2.3.5 Obecná chyba (5)

Pokud nastane obecná chyba aplikace, je vrácen kód 5.

5.2.3.6 Interní chyba (99)

V případě, že nastane neočekávaná interní chyba aplikace je vrácen kód 99.

5.3 IMPLEMENTAČNÍ DETAILY

Program je napsaný v jazyce C++ za použití knihoven OpenSSL pro práci s TLS a pugixml pro parsování formátu XML. Překlad probíhá pomocí CMake volaný přes Makefile. Překlad lze spustit pomocí příkazu `make`.

5.3.1 Návrh

Díky použití C++ je program navržen objektově. Rozdělen je do několika hlavních objektů: SSLWrapper (5.3.1.1), Http (5.3.1.2), Feed (5.3.1.4) a Url (5.3.1.3).

5.3.1.1 SSLWrapper

Tato třída abstrahuje volání knihovny OpenSSL a zjednodušuje práci s ní. Využívá BIO sokety pro nezabezpečené i zabezpečené připojení. Třída má nastavený časový limit (timeout) požadavků na 10 sekund.

5.3.1.2 Http

Třída Http abstrahuje vytváření a přijímání HTTP požadavků. K připojení k serveru využívá třídu SSLWrapper. Třída má jedinou veřejnou statickou metodu `get_request`, která pošle požadavek GET na zadanou URL adresu a vrátí obsah odpovědi serveru. V případě, že server vrátí chybu, tak tato metoda vyhodí výjimku.

5.3.1.3 Url

Třída Url je pomocná třída, která zajišťuje parsování URL na jednotlivé částí: protokol, doména, port a cesta. Tuto třídu pak využívá třída Http pro sestavení požadavku a SSLWrapper pro připojení k danému serveru.

5.3.1.4 Feed

Tato třída parsuje z XML informace pro zobrazení uživateli za použití knihovny pugixml. Parsuje se titulek zdroje a jednotlivé položky zdroje, u kterých se získává jejich titulek, autor, odkaz a čas poslední aktualizace. Toto je blíže popsáno v sekci 5.3.2.

5.3.2 Parsování zdroje

Parsování XML ze zdroje je uděláno stylem „best effort“, kde parser nehledí na formát zdroje, který zpracovává, ale hledá v XML elementy ze všech formátu, které podporuje. Elementy hledá podle specifikace, tak jak je uvedeno v sekcích 3.1 a 3.2, akorát povoluje absenci povinných elementů.

Parser implementuje i jednoduchou podporu jmenných prostorů pro RDF, Dublin Core a Atom. Deklarace jmenného prostoru musí být uvedena v kořenovém elementů. U RDF a Dublin Core povoluje taky absenci deklarace jmenného prostoru při použití standardních prefixů (rdf, dc).

5.3.3 Testování

Pro testování aplikace je napsán jednoduchý shell skript `test.sh`, který rekurzivně najde všechny soubory s příponou .test ve složce test/. Tyto soubory obsahují informace o jednotlivých testech

(jméno, parametry pro spuštění, návratový kód a jestli je třeba kontrolovat výstup aplikace). Pokud je nastavena kontrola výstupu, skript zkontroluje odlišnost výstupu programu oproti stejnojmennému souboru s příponou .out.

Testy lze spustit pomocí příkazu `make test`.

6 REFERENCE

1. **Přispěvatelé Wikipedie.** RSS. *Wikipedie.* [Online] Wikipedie: Otevřená encyklopedie, 4. 10. 2017. [Citace: 17. 11. 2018.] https://cs.wikipedia.org/w/index.php?title=RSS&oldid=15394320.

2. **UserLand Software.** RSS 0.91 Specification. *RSS Advisory Board.* [Online] 9. 6. 2000. [Citace: 18. 11. 2018.] http://www.rssboard.org/rss-0-9-1 (anglicky).

3. —. RSS 2.0 Specification. *RSS Advisory Board.* [Online] 14. 7. 2003. [Citace: 18. 11. 2018.] http://www.rssboard.org/rss-2-0 (anglicky).

4. **RSS-DEV Working Group.** RDF Site Summary (RSS) 1.0. *web.resource.org.* [Online] 9. 6. 2008. [Citace: 18. 11. 2018.] http://web.resource.org/rss/1.0/spec (anglicky).

5. **Easuwaran, Sathish.** RSS vs Atom. *Saksoft.* [Online] Saksoft Limited, 7. 11. 2015. [Citace: 18. 11. 2018.] https://www.saksoft.com/rss-vs-atom/ (anglicky).

6. **Nottingham, M. a Sayre, R.** RFC 4287 - The Atom Syndication Format. *IETF Tools.* [Online] 17. 12. 2005. [Citace: 18. 11. 2018.] https://tools.ietf.org/html/rfc4287 (anglicky).

7. **Rescorla, E.** RFC 8446 - The Transport Layer Security (TLS) Protocol Version 1.3. *IEFT Tools.* [Online] 10. 8. 2018. [Citace: 18. 11. 2018.] https://tools.ietf.org/html/rfc8446 (anglicky).