Filtr protokołu SMTP

Praca inżynierska

Zbigniew Artemiuk
Z.Artemiuk@stud.elka.pw.edu.pl

16 lipca 2008

Spis treści

1 Wstęp

Obecnie w Internecie, pomimo szeregu rozwijających się form komunikacji, chociażby technologii takich jak tekstowa komunikacja czasu rzeczywistego (czyli IM - instant messaging i popularne chaty) czy też jeszcze bardziej zaawansowanych technologii, które umożliwiają przesyłanie głosu oraz wideo (chociażby Skype), pospolite maile cały czas znajdują zastosowanie. Używamy ich chyba obecnie najczęściej do kontaktów biznesowych, rodzinnych, przyjacielskich, głównie w celach informacyjnych (czyli sam tekst), jak również do przekazywania niewielkich plików. Chyba każdy internauta legitymuje się przynajmniej jedną skrzynką mailową (a czasami jest ich znacznie więcej). Każdy spotkał się także z niechcianą pocztą (niechcianymi mailami), nazywaną bardzo ogólnym określeniem Spam, i narzędziami (chociażby tymi wbudowanymi w klientów pocztowych dostępnych z poziomu przeglądarki internetowej), które umożliwiają odseparowanie takiej poczty, od tej która niesie interesującą nas zawartość merytoryczną. Separacja ta to tzw. filtrowanie ze względu na obiekt filtracji zwane filtracja poczty. Dokonanie filtracji wymaga jednak poznania dokładnie w jaki sposób transportowane są nasze maile, a to wszystko zawarte jest w protokole Simple Mail Transfer Protocol (w skrócie SMTP). Coś tutaj jeszcze dopisać o protokole, o jego wadach, o konieczności filtracji.

2 Cel i zakres pracy

Celem pracy jest zaprojektowanie oprogramowania, które po uruchomieniu umożliwiałoby filtrację wiadomości w ramach protokołu SMTP. Oprogramowanie to byłoby konfigurowalne poprzez plik tekstowy zawierający reguły dotyczące wiadomości SMTP i zachowania oprogramowania po napotkaniu takiej wiadomości. Reguły te dawałyby możliwość następującej filtracji:

- wykrywanie wiadomości posiadających bądź nie posiadających określony nagłówek (nagłówki)
- wykrywanie wiadomości posiadających bądź nie posiadających określonej frazy w nagłówku (nagłówkach)
- wykrywanie wiadomości posiadających bądź nie posiadających określonej frazy w nagłówkach części maila (gdy nie jest to prosty mail)
- wykrywanie określonych content-type-ow w wiadomościach
- wykrywanie wiadomości o określonej wielkości części (jeżeli wiadomość jest jednoczęściowa to tyczy się treści wiadomości)

Po natrafieniu na wiadomość spełniającą kryteria danej reguły zostanie ona w zależności od zapisu w konfiguracji:

- odrzucona (wyfiltrowana)
- przepuszczona, a jej obecność zostanie zaznaczona w logach
- przepuszczona

3 Protokół SMTP

3.1 Historia powstania

3.1.1 ARPANET czyli początki Internetu

Historia powstania protokołu SMTP jest ściśle związana z początkami Internetu. Internet zaś i jego kreowanie związane jest bezpośrednio ze swoim przodkiem czyli ARPANETEm.

ARPANET (Advanced Research Projects Agency Network) został stworzony przez przez jedną z agencji United States Department of Defense (departament bezpieczeństwa Stanów Zdjednoczonych) o nazwie ARPA (Advanced Research Projects Agency). Nazwa agencji została później przekształcona na DARPA (D od Defence). Agencja ta miała zając się rozwojem nowych technologii na potrzebę amerykańskiego wojska.

W miarę wchodzenia w życie komputerów wykorzystywanych w ramach agencji powstała idea stworzenia sieci pomiędzy nimi, która to umożliwiłaby komunikacje pomiędzy ich użytkownikami. Idea ta została po raz pierwszy zaproponowana przez Josepha Carla Robnetta Licklidera z firmy Bolt, Beranek and Newman (obecnie BBN Technologies) w sierpniu 1962 w serii notatek na temat koncepcji "Międzygalaktycznej Sieci Komputerowej". Zawierała ona prawie wszystko czego możemy doświadczyć w dzisiejszym Internecie.

W październiku 1963 roku Licklider został mianowany szefem programu Behavioral Sciences and Command and Control w ARPA. Przekonał on wtedy Ivana Sutherlanda i Boba Taylora, że jego wizja jest czymś naprawdę istotnym. Sam Licklider nie doczekał jednak żadnych konkretnych prac w kierunku jej urzeczywistnienia, gdyż opuścił ARPA.

ARPA i Taylor cały czas byli zainteresowanie stworzeniem sieci komputerowej, ażeby zapewnić naukowcom pracującym w ramach ARPA w różnych lokalizacjach, dostęp do innych komputerów, które firma oferowała. Istotne było także, aby nowe oprogramowanie i rezultaty badań były jak najszybciej widoczne dla każdego użytkownika sieci. Sam Taylor posiadał 3 różne terminale, które dawały mu podłączenie do 3 różnych komputerów - jeden do SDC Q-32 w Santa Monica, drugi w ramach projektu Project Genie do komputera na Uniwersytecie w Kalifornii (Berkley) i ostatni do komputera z Multicsem w MIT (The Massachusetts Institute of Technology).

Taylor w taki sposób opowiadał od podłączeniu do tych komputerów: " Dla każdego z tych terminali miałem inny zestaw poleceń. Dlatego też kiedy rozmawiałem z kimś z Santa Monica, a poźniej chciałem ten sam temat skonsultować z kimś z Berkley albo MIT, musiałem przesiąść się do innego terminala. Oczywistym wtedy wydało mi się, że musi być 1 terminal, który obsłuży te 3 połączenia. Idea ta to właśnie ARPANET"

Do połowy 1968 roku kompletny plan sieci został stworzony i po zatwierdzeniu przez ARPA, zapytanie ofertowe RFQ (Request For Quotation) zostało posłane do 140 potencjalnych wykonawców. Większość potraktowała propozycję jako dziwaczną. Tylko 12 firm złożyło oferty z czego 4 zostały uznane za najważniejsze. Do końca roku wyłoniono 2 firmy, z których ostatecznie 7 kwietnia 1969 roku została wybrana firma BBN.

Propozycja BBN była najbliższa planom ARPA. Pomysłem ich było stworzenie sieci z mały komputerów zwanych Interface Message Processors (bardziej

znanych jako IMPs), które to obecnie nazywamy routerami. IMPsy z każdej strony zapewniały funkcje przechowywania i przekazywania pakietów, a połączone były między sobą przy użyciu modemów podpiętych do łączy dzierżawionych (o przepustowości 50 kbit/sekundę). Komputery podłączone były do IMPsów poprzez specjalny bitowy interfejs. W ten sposób stawały się one częścią sieci ARPANET.

Do zbudowania pierwszej generacji IMPsów BBN wykorzystała komputer Honeywell DDP-516. Został on wyposażony w 24kB pamięci rdzenia (z możliwością rozszerzenia) oraz 16 kanałów Direct Multiplex Control (DMC) do bezpośredniogo dostępu do tej pamięci. Poprzez DMC podłączane były komputery użytkowników (hosty) i modemy. Dodatkowo 516 otrzymał ezstaw 24 lamp, które pokazywały status kanałów komunikacyjnych IMPa. Do każdego IMPa można było podłączyć do czterech hostów i mógł się on komunikować z 6 zdalnymi IMPami poprzez współdzielone łącza.

Zespół z BBN (początkowo 7 osób) szybko stworzył pierwsze działające jednoski (IMPy). Cały system, który zawierał zarówno sprzęt jak i pierwsze oprogramowania zarządzające pakietami, został zaprojektowany i zainstalowany w ciągu 9 miesięcy.

Początkowo ARPANET składał się z 4 IMPów. Zostały one zainstalowane w:

- UCLA (University of California, Los Angeles), gdzie Leonard Kleinrock założył centrum pomiaru sieci (Network Measurment Center)
- The Stanford Research Institute's Augmentation Research Center, gdzie Douglas Engelbert stworzył system NLS, który między innymi wprowadził pojęcie hypertextu
- UC Santa Barbara
- The University of Utah's Graphics Department, gdzie przebywał ówcześnie Ivan Sutherland

3.1.2 Piersze wiadomości w ARPANETcie

Pierwsza komunikacja host-host w sieci ARPANET wykorzystywała protokół 1822, który definiował sposób w jakis host przesyłał wiadomość do IMPa. Format wiadomości był tak zaprojektowany, żeby bez problemu mógł pracować z szerokim zakresem architektur. Zasadniczo wiadomość składała się z:

- typu wiadomości
- adresu hosta
- pola z danymi

W celu wysłania wiadomości do innego hosta, host wysyłający powinien sformatować wiadomość tak, aby ta zawierała adres hosta docelowego oraz dane, a następnie dokonać transmisji wiadomości przez interfejs sprzętowy 1822. IMP dostrzeże dostarczenie wiadomości albo poprzez dostarczenie jej bezpośrednio do hosta docelowego albo poprzez przekazanie jej do kolejnego IMPa. Kiedy wiadomość została odebrana przez docelowego hosta, IMP do którego host był

podłączony wysyła potwierdzenie odbioru (zwane Ready for Next Message or RFNM) do hosta wysyłającego.

W przeciwieństwie do obecnych protokołów datagramowych w Interecie (takich jak no IP), ARPANETowy porotokół 1822 zapewniał niezawodność w taki sposób, że informował o niedostarczonej wiadomości. Niemniej protokół 1822 nie był odpowiedni do żonglowania wieloma połączeniami w różnych aplikacjach uruchomionych na pojedynczym hoście. Problem ten został rozwiązany dzięki wprowadzeniu na hostach Network Control Program (NCP), dzięki któremu możliwe było niezawodne, z kontrołą przepływu, dwukierunkowe połączenia pomiędzy różnymi procesami na różnych hostach. NCP implementował kolejną warstwę znajdującą się na górze stosu protokołów. Dzięki niemu aplikacje, które miały mieć już jakąś konkretną funkcjonalność, mogły wykorzystywać spójny interfejs i korzystać swobodnie z dobrodziejstw ARPANETu czyli wykonywać połączenia do innych aplikacji przez sieć.

3.1.3 SNDMSG i READMAIL

Na początku roku 1970, powstał program (a w zasadzie 2 oddzielne) do wysyłania i odbierania wiadomości. Zaimplementował go Ray Tomlinson. Programy te to SNDMSG i READMAIL. Pierwsza wersja tych programów, była kolejną implementacją, która umożliwiała wymiane informacji między użytkownikami jednej maszyny (jednego hosta), a konkretnie komponowanie, adresowanie i wysyłanie wiadomości do skrzynek użytkowników. Już jednak w 1971 Tomlinson stworzył pierwszą aplikację ARPANETową (w ramach prac nad systemem TENEX), która umożliwiała wysyłanie wiadomości do dowolnych hostów. Tomlinson dokonał usprawnień w programie SNDMSG przy okazji pracy nad projektem CPYNET, którego celem było stworzenie protokołu do wymiany plików między komputerami w sieci.

Skrzynkę emailową był wówczas był zwykły plik o określonej nazwie. Specjalną właściwością jego było to, że miał ochronę taką, że umożliwiał innym użytkownikom dopisywanie do pliku. Mogli oni więc pozostawiać kolejne wiadomości ale nie mogli ich czytać ani nadpisywać wcześniejszych (jedynie właściciel mógł to robić). Tomilson zauważył, że CPYNET może dopisywać zawartość do pliku skrzynki mailowej, na takiej zasadzie jak SNDMSG (SNDMSG umiał zapisywać wtedy tylko na lokalnej maszynie). Dlatego też SNDMSG mógł po prostu skorzystać z kodu CPYNET i bezpośrednio dostarczać wiadomości poprzez połączenie sieciowe do zdalnych skrzynek poprzez dopisywanie kolejnych informacji do plików na innych hostach.

Brakującą częścią była możliwość dopisywania do plików przy pomocy CPYNET. Dotychczas mógł on tylko słać i odbierać pliki. Dodanie tej funkcjonalności nie było czymś wielkim dla twórców protokołu i funkcjonalność ta wkrótce zainstniała.

Następnie Tomilson włączył kod CPYNET do SNDMSG. Pozostało jedynie rozróżnienie maili lokalnych od maili zdalnych. Dlatego też Tomilson zdecydował się, że maile zdalne będą rozpoznawane po tym, że po loginie (czyli wyznaczniku użytkownika do którego wiadomość jest słana) nastąpi specjalny znaczek @ (ang. at, a polska znana wszystkim małpa) , a tuż po nim nazwa hosta czyli zdalnego komputera, na którym skrzynkę ma dany loginem użytkownik. Tomilson tak powiedział o wyborze małpy jako znaczka rozdzielającego login od hosta: "I am frequently asked why I chose the at sign, but the at sign just makes

sense" (w wolnym tłumaczeniu: Jestem często pytany czemu wybrałem znaczek małpy, ale on po prostu miał sens).

Pierwsza wiadomość została przesłana pomiędzy maszynami, które fizycznie były obok siebie. Jedyne połączenie jakie było między maszynami (oprócz podłogi ;]) było za pośrednictwem sieci ARPANET. Tomilson przesłał wiele wiadomości do samego siebie z jednej maszyny na drugą. Pierwsze treści wiadomości od razu zostały zapomniane. Najprawdopodobniejsza ich treść była w stylu QWERTYUIOP. Kiedy Tomilson był zadowolony z programu wysłał do swoich kolegów z zespołu wiadomość z instrukcją jak słać wiadomości przez sieć. I tak pierwsza wysłana wiadomość ogłosiła swoje istnienie.

Te pierwsze wiadomości zostały wysłane pod koniec 1971 roku. Następne wydanie TENEXa, które ukazało się w 1972 roku, zawierało SNDMSG, z możliwością wysyłania maili przy użyciu sieci ARPANET. Protokół CPYNET wkrótce został zastąpiony prawdziwym protokołem do transferu plików i posiadał specyficzne dodatki do obsługi maila.

3.1.4 MAIL i MLFL w protokole FTP

Protokołem, o którym wcześniej była mowa był protokół FTP, którego specyfikacja wstępna została zawarta w RFC (Request For Comment) 114. Kolejne ulepszenia protokołu zostały dokonana w roku 1972 i wtedy też weszły do niego nowe polecenia pozwalające korzystać ze skrzynek emailowych. Poleceniami tymi były:

- MAIL - polecenie to pozwalało użytkownikowi przy pomocy telnetu wysłać maila. Pomocne to było gdy użytkownik ten nie korzystał ze swojego hosta i logował sie do niego poprzez protokól telnet
- MLFL - polecenie to pozwalało na normalne skonstruowanie maila i przesłanie odpowiednio wskazanego pliku jako jego treści

Protokół ten stał się aż do roku 1980 standardem przesyłania emaili w ARPANETcie. Wtedy to został wypart przez protokół SMTP, który z pewnymi usprawnieniami funkcjonuje aż do dziś.

3.1.5 Rozwój innych programów do obsługi poczty

Zanim jednak powstał protokół SMTP cały czas pracowano nad rozwojem ówczesnych programów chociażby do odbioru maili. W ten sposób na prośbe Steve'a Lukasika Lawrence Roberts, ówczesny dyrektor IPTO, stworzył program RD, który składał się z makr w edytorze TECO (Text Editor and COrrector).

Program RD umożliwiał:

- sortowanie emaili po temacie i dacie
- czytanie, zapisywanie i czytanie wiadomości w dowolnym porządku

RD nie powstał więc jako wynik badań, ale z czysto praktycznej potrzeby swobodnego zarządzania emailami.

Wkrótce powstały też kolejne ulepszenia programu RD oraz SENDMSG, takie jak NRD czy WRD, które to wprowadzały kolejne ulepszenia istniejących już implementacji.

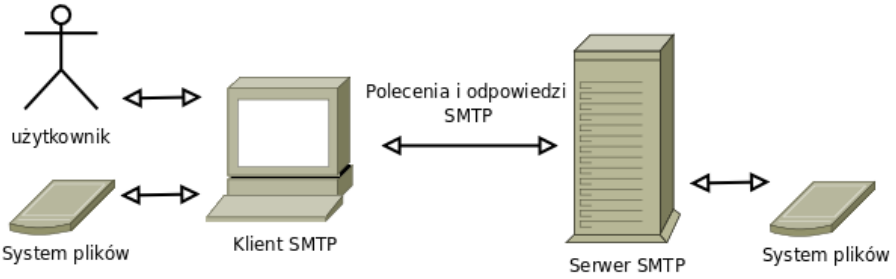
Warte wspomnienia jest także powstanie programu MSG, który umożliwiał między innymi przekazywanie maili (ang. forwarding), konfigurowalny interfejs, czy też polecenie Odpowiedz (ang. Answer), które automatycznie uzupełniało adres odbiorcy. MSG można było nazwać pierwszym nowoczesnym programem pocztowym.

W 1977 roku różne formaty wiadomości zostały zebrane w jedną spójną specyfikację co doprowadziło do stworzenia RFC 733. Specyfikacja ta połączyła wcześniej istniejące dokumentację z odrobioną innowacji i była jednocześnie pierwszym RFC, które deklarowało standard Internetowy (ówcześnie ARPANETowy).

3.2 Szczegóły protokołu

W 1982 roku powstało główne RFC opisujące protokół SMTP - RFC 821. W 2001 roku zostało ono rozbudowane i powstał dokument, który jest obowiązującą obecnie specyfikacją protokołu. Informacje te zawarte zostały w RFC o numerze 2821.

3.2.1 Model protokołu, podstawy



Komunikacja w protokole SMTP oparta jest na nastepującym modelu: kiedy klient SMTP posiada wiadomość, którą chce przetransmitować ustanawia dwukierunkowy kanał transmisyjny z serwerem SMTP. Klient SMTP bierze na siebie odpowiedzialność za przetransportowanie wiadomości do jednego lub wielu serwerów SMTP, lub zgłoszenie błędu jeśli operacji przekazania maila nie powiodła się.

Adres serwera SMTP klient determinuje poprzez zamianę domeny podanej w adresie na pośredni host (Mail eXchanger) lub ostateczny host docelowy.

Serwer SMTP może być ostatecznym celem albo też pośrednim "przekaźnikiem" (po odebraniu wiadomości to on może przejąć rolę klienta SMTP) albo "bramą" (może transportować wiadomość dalej używając innego protokołu niż SMTP). Polecenia protokołu SMTP generowane są przez klienta i posyłane do serwera. Odpowiedzi słane są jak reakcja na żądania klienta.

Innymi słowy wiadomość może być przetransportowana w trakcie pojedynczego połączenia pomiędzy pierwotnym adresatem, a ostatecznym klientem albo może składać się z kilku pośrednich połączeń. W obu przypadkach zachowana jest odpowiedzialność za wiadomość - protokół wymaga od serwera wzięcia odpowiedzialności za dostarczenie wiadomości albo w przypadku nieprawidłowości za zgłoszenie błędu.

Kiedy kanał transmisyjny zostanie zestawiony klient SMTP inicjuje transakcję maila. Składa się ona z serii poleceń, które mają na celu wyspecyfikowanie

nadającego i celu wiadomości, a następnie przesłania jej zawartości (oczywiście z nagłówkami, które wchodzą w jej skład). Kiedy ta sama wiadomość jest adresowana do wielu odbiorców, protokół przesyła jedną kopię wiadomość dla wszystkich odbiorców w obrębie jednego hosta.

Serwer reaguję na każdą komendę odpowiedziami. Wskazują one akceptację komendy (tej przesłanej od klienta) i oczekiwanie na kolejne, bądź też wystąpienie tymczasowych albo stałych błędów.

Kiedy mail zostanie przetransmitowany, klient może zamknąć połączniee albo też zainicjować kolejną tranzakcję innego maila. Dodatkowo klienty SMTP mogą używać połączenia z serwerem w celach np. weryfikacji adresu email bądź też uzyskanie adresów subskrybentów listy dyskusyjnej.

Jak zostało wcześniej zasugerowane transmisja w protokole może odbywać się bezpośrednio pomiędzy hostem ślącym wiadomośc, a hostem docelowym, kiedy są one połączone tym samym serwisem transportowym. Kiedy tak nie jest, transmisja odbywa się poprzez hosty pośrednie. Hosty te wybierane są poprzez mechanizm serwera domen (DNS) zwany Mail eXchanger.

3.2.2 Model rozszerzeń

W wyniku prac, które rozpoczęły się około dekadę po wydaniu pierwszego RFC dotyczącego protokołu SMTP (czyli RFC 822), protokół został rozszerzony i pojawiły się w nim nowe funckjonalności. Standardowy model został więc zmodyfikowany o dodatkowe serwisy. Pozwalają one na ustalenie pomiędzy klientem, a serwerem usług dodatkowych jakie są oni w stanie obsłużyć (poza podstawowymi wymaganiami SMTP). Mechanizm rozszerzeń SMTP określa środki, za pomocą których klient i serwer mogą dokonać wzajemnego ropoznania, a serwer może także poinformować klienta o rozszerzeniach, które on wspiera.

Współczesna implementacja SMTP musi obsługiwać podstawowe mechniazmy rozszerzeń. Przykładowo serwer musi wspierać komendę EHLO, jeśli nawet nie implementuje innych specyficznych rozszerzeń, natomiast klient powinien skłaniać się ku używaniu komendy EHLO niż HELO.

3.2.3 Konstrukcja wiadomości

Konstrukcja

3.3 Obecne wykorzystanie protokołu i jego forma

Obecne wykorzystanie

4 Dzisiejsze narzędzia do filtracji protokołu SMTP

Dzisiejsze

4.1 Konieczność wprowadzenia filtracji

Konieczność

4.2 Produkty komercyjne

Produkty komercyjne

4.2.1 Clearswift

Clearswift

4.2.2 Aladdin eSafe

Aladdin eSafe

4.2.3 Surfcontrol Email Filter

Suontrol Email Filter

4.3 Produkty open-source

5 Opracowany filtr poczty SMTP

5.1 Założenia projektu

5.2 Moduły projektu

5.2.1 Parser wiadomości

5.2.2 Parser reguł

5.2.3 Analizator wiadomości

5.2.4 Kolejka

5.3 Kompilacja, konfiguracja i uruchomienie projektu

5.4 Testy wydajnościowe

6 Spostrzeżenia, wnioski

Literatura

[Cro82] David H. Crocker. *RFC822 - STANDARD FOR THE FORMAT OF ARPA INTERNET TEXT MESSAGES*. 1982.