Scientific Bulletin of Chełm
Section of Mathematics and
Computer Science
No. 1/2008

INFOBOT.PL - BOT INFORMACYJNY

FILIP KWIATKOWSKI, MACIEJ SZEWCZYK

Uniwersytet Marii Curie-Skłodowskiej w Lubinie

Streszczenie. Artykuł prezentuje Infobota - najpopularniejszego bota informacyjnego w Polsce. Przedstawione informacje opisują sposób działania bota, agregacji danych i udostępniania ich użytkownikom. W pracy skupiliśmy się przede wszystkim na problemach, z jakimi można się spotkać tworząc podobne rozwiązania i sposobach, w jaki można je rozwiązać.

"Wszystko powinno zostać uproszczone tak bardzo,
jak to tylko możliwe, ale nie bardziej."
Albert Einstein

1. Czym jest Infobot?

Bot to program, który w sposób automatyczny wykonuje pewne często powtarzane czynności. Jednym z rodzajów botów są boty informacyjne, takie jak Infobot [1]. Oferuje on blisko 30 funkcji, a wśród nich: słowniki językowe (takie jak np. angielsko-polski), słownik języka polskiego, ortograficzny, wyrazów bliskoznacznych. Poza tym: program telewizyjny, prognozę pogody, wyniki Lotto, dostęp do Wikipedii, daty imienin i skracacz adresów internetowych. Dostęp do tej funkcjonalności można uzyskać w sieciach Gadu-Gadu [2], Skype [3], Tlen [4], Jabber [5], a także przez pocztę Google Mail [6] i SMSy. Wystarczy wysłać do Infobota krótkie zapytanie, aby po chwili dostać odpowiedź, co ukazuje zrzut ekranu (rysunek 1).

2. Historia projektu

Historia projektu zaczęła się w 2004 roku, gdy doszliśmy do wniosku, że wygodnie byłoby mieć dostęp do prognozy pogody z poziomu komunikatora internetowego. Znalezienie prognozy w Internecie wiąże się z wieloma czynnościami, które trzeba za każdym razem powtarzać. Poza tym, oprócz tych danych, które nas interesują, musimy pobierać wiele innych, nadmiarowych, jak na przykład elementy graficzne czy reklamy. Podobnie

Treść artykułu była prezentowana w czasie *VII Konferencji Informatyki Stosowanej* (Chełm 30 - 31 maja 2008 r.)

ma się sprawa z innego rodzaju informacjami, takimi jak słowniki. Nasze rozwiązanie okazało się wygodne nie tylko dla nas. Obecnie z Infobota skorzystało już ponad milion polskich internautów.

Rysunek 1

3. Schemat działania

Na samym początku Infobota tworzyło kilkanaście prostych skryptów, napisanych w języku BASH [7], natomiast za połączenie z siecią Gadu-Gadu odpowiadał eksperymentalny klient gadu gadu [8] (ekg), dostępny na platformy uniksowe. Z czasem jednak przestało to być rozwiązaniem wystarczającym. Coraz większa popularność Infobota i związane z tym obciążenie zmusiły nas do napisania Infobota od podstaw. Wybraliśmy język C, a to ze względu na jego dużą wydajność, związaną głównie z istnieniem bardzo dobrych kompilatorów, które potrafią w wysokim stopniu zoptymalizować wynikowy plik binarny. Istnieją oczywiście języki, których działanie jest szybsze od skryptów BASH, a jednocześnie pozwalają na wydajne pisanie programów. Przykładem takiego języka jest Python [9], który posiada między innymi automatyczne zarządzanie pamięcią. C jest pod wieloma względami mniej wygodne i trzeba poświęcić więcej czasu na napisanie takiego samego programu, jednakże - według testów - nie ma sobie równych, jeśli chodzi o zużycie pamięci i szybkość działania [10]. Do łączenia się Infobota z odpowiednimi sieciami wykorzystujemy biblioteki, takie jak libgadu (Gadu-Gadu), libtlen (Tlen), loudmouth (Jabber).

4. Bazy danych

Ponieważ podstawą działania Infobota jest oferowanie informacji, siłą rzeczy korzystamy z baz danych do ich przechowywania. Tu, podobnie jak przy wyborze języka, kierowaliśmy się jak najlepszą wydajnością. Początkowo stosowaliśmy bazę danych SQLite [11], która jest mała i szybka, jednakże jedną z jej cech jest brak centralnego procesu, odpowiadającego za obsługę poleceń. Sprawia to, że nie zawsze radzi sobie z problemem konkurencyjności. Chwilowo wykorzystywaliśmy również bazę Oracle BerkeleyDB [12], będącą bazą dość prostą, bo nie relacyjną, ale jednocześnie niezwykle szybką. Baza ta przechowuje dane w relacji klucz-wartość. Z czasem okazało się jednak, że słaba obsługa konkurencyjności przez te bazy wyklucza ich użycie w coraz bardziej obciążonym Infobocie. W związku z tym postanowiliśmy porównać popularne systemy MySQL [13] i PostgreSQL [14]. Najpierw sprawdziliśmy działanie tych systemów bazodanowych w wersjach, odpowiednio, 8.0.15 i 5.0.44. Test polegał na dodaniu do bazy 10 milionów wierszy, założeniu indeksów typu b-tree oraz hash i porównaniu czasu wykonania operacji takich jak insert, delete oraz select przy relacjach "=" oraz "<". Wyniki były dość zaskakujące, wskazując na dużą przewagę MySQL. Dość "podejrzane" były jednak czasy zapytań select, zawierające relację "<", w zasadzie nie różniące się między sobą, bez względu na to, czy zastosowano indeksowanie, czy też nie. Dlatego dokonaliśmy również porównania z PostgreSQL w wersji 8.2.6. Wyniki przedstawione zostały w poniższej tabeli.

DBMS[1]	**indeks**	**create table**	**create index**	**insert**	**delete**	**select** (=)	**select** (<)
PostgreSQL 8.0.15	-	92 m	-	0,0001 s	4,5 s	4,5 s	4,5 s
	btree	-	1 m	0,0008 s	0,0008 s	0,0003 s	4,5 s
	hash	-	7 h	0,0008 s	0,0008 s	0,0003 s	4,5 s
MySQL 5.0.44	-	11 m	-	0 s	4,1 s	3,8 s	4 s
	btree	-	78 s	0 s	0,03 s	0,03 s	0,03 s
PostgreSQL 8.2.6	-	92 m	-	0,0007 s	2,8 s	2,8 s	2,8 s
	btree	-	46 s	0,0008 s	0,0008 s	0,0003 s	0,0003 s
	hash	-	10 h	0,0008 s	0,0008 s	0,0003 s	2,8 s

Jak widać, w wersji 8.0.15 PostgreSQL, w przypadku zapytań zawierających relację "<", indeksowanie nie dawało żadnego przyspieszenia działania. Było to do przewidzenia w indeksie typu hash, gdyż funkcje haszujące pozwalają szybko znaleźć konkretne dane pod kluczem, jaki dokładnie znamy. Natomiast indeksowanie typu b-tree powinno ponadto przyspieszać wyszukiwania z relacjami mniejszości i większości. Okazało się, że w wersji 8.2.x to dziwne zachowanie zostało naprawione. Widać również, że w przypadku MySQL szybszy był czas dodania danych do bazy. Było to jednak okupione dużo wyższym użyciem czasu procesora. Biorąc pod uwagę wszystkie "za i przeciw", zdecydowaliśmy się na użycie PostgreSQL. Dużą zaletą tego systemu jest również wygodna konsola, z jakiej można zarządzać bazami.

[1] *Database Management System* - System Zarządzania Bazą Danych

Poza przechowywaniem informacji, bazy danych wykorzystujemy również do zapobiegania atakom typu (D)DOS [15]. Przechowujemy na bieżąco informacje o wiadomościach wysłanych do Infobota w przeciągu ostatnich 10 sekund. W przypadku, gdy jakiś użytkownik przekroczy dopuszczalną liczbę wysłanych wiadomości, zostaje zablokowany na kilka minut.

5. Infobot w sieci Skype

Infobot działający w sieci Skype wymaga uruchomionego oryginalnego klienta tej sieci, gdyż jest ona zamknięta i nikt do tej pory nie rozszyfrował jej protokołu. Producent tego komunikatora udostępnił jednakże API [16], czyli interfejs, jakim można się komunikować z programem. Umożliwia on reagowanie na każde zdarzenie, jakie pojawi się podczas działania programu. W systemie Linux można korzystać z API poprzez protokoły D-BUS [17] oraz X11 [18]. Ze względu na częste zmiany w tym pierwszym, prowadzące do niekompatybilności z wcześniej napisanymi programami i na dość stabilną specyfikację protokołu X11, do komunikacji ze Skype wybraliśmy właśnie ten drugi. Istnieją, co prawda, gotowe biblioteki, pośredniczące w komunikacji z API Skype'a, jednakże bezpośrednie odwoływanie się do API zapewnia większą elastyczność. Po włączeniu programu Skype prosi o zezwolenie na komunikację z tymże programem. Jeśli się zgodzimy, nasz program będzie odbierał zdarzenia od Skype'a, a także będzie mógł mu wysyłać polecenia, np. wysłania wiadomości. Infobot nie traktuje nadchodzących zapytań jako kolejki wiadomości i obsługuje je jako osobne procesy - głównie ze względu na to, że obsługa niektórych wiadomości nie odbywa się w sposób błyskawiczny, przez co późniejsze wiadomości musiałyby czekać na obsłużenie poprzednich. Okazało się jednak, że protokół X11 niezbyt dobrze radzi sobie z obsługą asynchronicznych komunikatów.

Dlatego zdecydowaliśmy się rozwiązać to nieco inaczej. Do komunikacji ze Skype tworzymy dwa procesy - pierwszy obsługuje wiadomości przychodzące, wynik działania zapisuje do bazy i powiadamia drugi proces, który odpowiada za wysyłanie wiadomości do użytkownika i usuwanie niepotrzebnych już wiadomości z bazy. Dzięki takiej niezależności obu procesów znika problem kłopotów z asynchronicznością otrzymywanych i wysyłanych komunikatów.

6. Jabber/XMPP

Jednym ze sposobów dostępu do Infobota jest sieć Jabber/XMPP. XMPP *(ang. Extensible Messaging and Presence Protocol)* to otwarty protokół, bazujący na języku XML, umożliwiający przesyłanie wiadomości oraz powiadamiający o obecności w czasie rzeczywistym. Dzięki zaangażowaniu w rozwój tego protokołu takich firm jak na przykład Google, HP czy Apple, istnieją wielkie szanse, że stanie się on głównym protokołem stosowanym tak w komunikatorach, jak i w innych systemach błyskawicznej komunikacji. Już teraz szacuje się, że za pomocą Jabbera porozumiewa się ze sobą około 90 milionów osób, a takie cechy, wyróżniające go spośród konkurencyjnych sieci, jak otwartość, decentralizacja, bezpieczeństwo oraz transporty, powodują, że zainteresowanie Jabberem ciągle rośnie. Decentralizacja - jedna z najważniejszych cech tej sieci - oznacza, że w Internecie dostępnych jest wiele serwerów Jabbera, a dodatkowo, że każdy może uruchomić własny serwer. My również skorzystaliśmy z takiej możliwości, udostępniając usługi Infobota pod identyfikatorem infobot@infobot.pl. Serwer, który zastosowaliśmy do obsłużenia tak wielkiej ilości wiadomości, to jabberd2 [19]. Dzięki skalowalności

i wsparciu dla najnowszych rozszerzeń protokołu XMPP, doskonale radzi on sobie z obsługą wiadomości przychodzących od ponad 250 serwerów Jabbera. W przeważającej mierze z naszych usług korzystają użytkownicy Google Talk/Gmail.

7. Google Mail

Bardzo liczne grono naszych użytkowników korzysta z usługi Google Mail. Pozwala ona na dodanie do książki adresowej kontaktów z sieci Jabber/XMPP. Oznacza to, że gdy dodamy do niej wspomniany wyżej identyfikator infobot@infobot.pl, będziemy mogli czatować z Infobotem przez interfejs WWW w oknie przeglądarki. Rozwiązanie to nie wymaga zainstalowanego komunikatora i sprawdza się doskonale np. podczas pisania e-maili.

8. Agregacja informacji

Informacje dostępne w Infobocie pobierane są z różnych stron WWW oraz - w przypadku funkcji "news" - z tzw. kanałów RSS. W związku z tym należy zwrócić uwagę na kilka nieuniknionych problemów. W przypadku stron WWW dane muszą zostać przeparsowane z kodu HTML do sformatowanego tekstu, gotowego do wysłania użytkownikom. Problemem, z jakim się tutaj spotykamy, jest to, iż strony internetowe używają różnych sposobów kodowania znaków narodowych. Podobnie też jest z sieciami, w których Infobot jest dostępny. W sieci Gadu-Gadu używane jest kodowanie CP-1250, co ogranicza możliwość wysyłania wielu znaków diakrytycznych w takich językach, jak na przykład francuski. W pozostałych sieciach, czyli Skype, Tlen oraz Jabber, problem ten nie istnieje, ponieważ używają one kodowania UTF-8. Poza tym, przy pobieraniu informacji ze stron trzeba brać pod uwagę zmiany ich wyglądu oraz czasową niedostępność. Warto napisać uniwersalny kod parsujący tak, aby nie był czuły na mniejsze zmiany.

Z problemami tymi w mniejszym stopniu spotykamy się przy pobieraniu danych dla funkcji "news". Infobot oferuje aktualne informacje z kraju, świata, gospodarki, nauki itp. W związku ze stałym napływem nowych wiadomości, najlepszym sposobem ich ciągłej agregacji jest korzystanie z czytnika RSS. *Really Simple Syndication* (RSS) to umowna rodzina języków znacznikowych, opartych na XML, które służą do przesyłania nagłówków wiadomości (tytułu i krótkiego opisu). Prawie każdy większy portal udostępnia streszczoną formę własnych wiadomości za pomocą kanałów RSS. W Infobocie czytnikiem odpowiedzialnym za pobieranie RSS-ów jest *Universal Feed Parser* [20], napisany w języku Python. Stała składnia dokumentów RSS pozwala na sprawne pobieranie informacji z wielu źródeł, a naszym użytkownikom na bycie na bieżąco z najświeższymi informacjami. Istotne jest również to, aby niektóre dane były pobierane regularnie. Za aktualność newsów, prognozy pogody, wyników lotto, kursów walut oraz programu tv odpowiedzialny jest program "cron". Jest to standardowe uniksowe narzędzie do zautomatyzowanego uruchamiania zadań o określonej porze.

9. Poprawianie błędów użytkowników

Podczas wpisywania poleceń użytkownikom zdarza się popełniać błędy - czy to przypadkiem, czy też nie do końca znając pisownię danego słowa. Infobot w przypadku nie znalezienia danego słowa w słowniku szuka podobnego słowa i sugeruje poprawną pisownię. Dotyczy to również innych funkcji, takich jak prognoza pogody.

Przykładowe błędy, z jakimi potrafi sobie poradzić, widoczne są poniżej:

- pf smaochód,
- pogdao Alksandów Kjujawksi,
- imn Brzenka,
- pa żyrandko;p,
- pn kucharz.

Na poniższym zrzucie ekranu widać, co Infobot robi w takiej sytuacji.

Rysunek 2

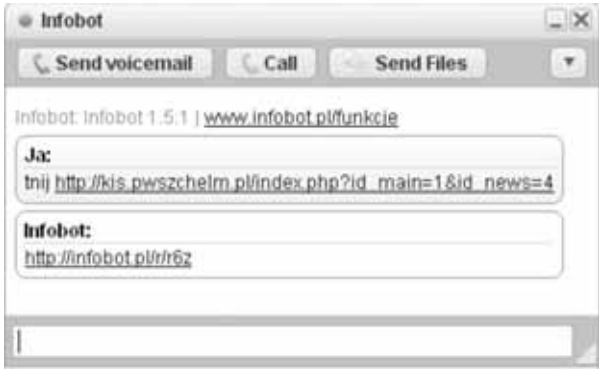
Poza literówkami radzi sobie również z błędami ortograficznymi. Do automatycznej korekcji błędów zastosowaliśmy szereg ciekawych algorytmów oraz wyrażenia regularne. Aby poprawić jeszcze bardziej trafność sugestii, chcemy dodatkowo brać pod uwagę odległość klawiszy na klawiaturze. Przykładowo: w przypadku błędnie wpisanego słowa "bryon", bardziej prawdopodobne jest, że chodziło o słowo "beton" niż "balon", mimo, że oba różnią się od słowa "bryon" dwoma znakami. Taka różnica między łańcuchami tekstowymi nazywana jest odległością Levenshteina i określa liczbę operacji, takich jak dodanie, usunięcie, czy zamiana znaków koniecznych do przekształcenia jednego łańcucha w drugi. Algorytmem, który implementuje tę metodę jest algorytm Wagnera-Fischera. Wykorzystuje on programowanie dynamiczne, w związku z czym jego złożoność obliczeniowa wynosi $O(mn)$, gdzie m i n są długościami obu łańcuchów. Dlatego też w pierwszym kroku stosujemy szybszy algorytm Wu-Manbera, który - mimo, że nie pozwala określić kosztów operacji, takich jak usunięcie czy wstawienie znaku - pozwala

na znalezienie podobnych łańcuchów. Następnie analizowany jest wynik jego działania, czyli lista najbardziej podobnych słów.

W przypadku wystąpienia błędu ortograficznego, dla danego słowa tworzone jest wyrażenie regularne poprzez zamianę głosek takich jak ż oraz rz na $[rz|ż|sz]$ - oznacza to, że na danej pozycji w tekście może wystąpić któraś z nich. Przykładowo, jeśli użytkownik wpisze "hżonżdż", powstanie wyrażenie $[h|ch]$ $[rz|ż|sz]$ $[ą|on|om]$ $[rz|ż|sz]$ $[dż|cz]$ i wśród listy prawidłowych słów zostanie odnalezione słowo "chrząszcz".

10. Skracacz adresów

Jedną z funkcji Infobota jest skracacz, który ma za zadanie skracanie długich adresów stron internetowych. Dostrzeżony przez nas problem długich linków, nie mieszczących się w opisach w ulubionych komunikatorach czy też SMS-ach oraz na grupach dyskusyjnych, postanowiliśmy rozwiązać, udostępniając naszym użytkownikom funkcję "tnij". Mechanizm skracania adresów internetowych jest w gruncie rzeczy dość prosty. Każdy zewnętrzny adres jest wiązany z adresem w domenie infobot.pl. Przykładowy adres tego typu wygląda tak: http://infobot.pl/r/r6z. Końcówkę stanowią kolejne cyfry w systemie sześćdziesięcio-dwójkowym, przy czym po cyfrach 0..9 kolejne liczby tego systemu zastępowane są symbolami małych i dużych liter. Pozwala to znacząco skrócić adres. Nawet przy miliardach zapisanych w bazie adresów końcówka ta będzie miała tylko kilka znaków. Przy dodawaniu kolejnego, skróconego adresu do bazy danych, numer nowego adresu przekształcany jest na wspomniany ciąg znaków klasyczną metodą, wykorzystującą dzielenie modulo i całkowite przez liczbę 62, dopóki liczba ta nie będzie równa 0.



Rysunek 3

Próba dodania po raz kolejny tej samej strony do bazy wyświetli - oczywiście - skrót już wcześniej utworzony. Zapobiega to duplikowaniu się skrótów w bazie. Obecnie posiadamy ponad 100 tysięcy aktywnych skrótów.

11. YouTube w Infobocie

YouTube jest popularnym serwisem internetowym, umożliwiającym publikację w Internecie filmów, wideoklipów czy też własnych produkcji. W ramach projektu Infobot uruchomiliśmy specjalną stronę http://youtube.infobot.pl/, na której umieszczane są

odnośniki do filmów z YouTube.com. Schemat działania strony jest bardzo prosty. Użytkownik przesyła do Infobota link do wideoklipu, który zostaje automatycznie umieszczony na naszej stronie w postaci miniaturki opatrzonej tytułem, czasem trwania oraz liczbą obejrzeń. Informacje te pobierane są za pomocą interfejsu programowania aplikacji (API), udostępnionego przez YouTube. Interfejs ten opiera się na Google Data API (GData). Jest to nowy protokół, oparty o XML, przypominający wspomniany wcześniej w artykule RSS. Program (napisany również w Pythonie) odpowiedzialny za pobieranie GData generuje następnie podstrony z wideoklipami rozmieszczonymi w postaci listy, siatki oraz mozaiki, które można sortować według popularności, czasu trwania lub dodania. Dzięki tej stronie użytkownicy mogą dzielić się swoimi propozycjami filmów wartych obejrzenia, a także sprawdzać, co jest obecnie chętnie oglądane.

12. Podsumowanie

Komunikatory internetowe są obecnie powszechnym środkiem komunikacji i pozwalają na błyskawiczne przesyłanie wiadomości. Infobot skrócił dystans pomiędzy użytkownikiem a informacjami, jakich w danym momencie potrzebuje, oferując bezpośredni do nich dostęp, bez konieczności żmudnej selekcji danych na stronach internetowych. Pozwala to zaoszczędzić sporo czasu. Do tej pory z Infobota skorzystało ponad milion osób i napisały one przeszło 160 milionów zapytań. Świadczy to o tym, że sposób dostępu do informacji, jaki oferuje Infobot, jest wygodny dla dużej grupy internautów.

Bibliografia

[1] http://www.infobot.pl/
[2] http://www.gadu-gadu.pl/
[3] http://www.skype.com/
[4] http://www.tlen.pl/
[5] http://www.jabber.org/
[6] http://www.gmail.com/
[7] http://www.gnu.org/software/bash/
[8] http://ekg.chmurka.net/
[9] http://www.python.org/
[10] http://shootout.alioth.debian.org
[11] http://www.sqlite.org/
[12] http://www.oracle.com/database/berkeley-db.html
[13] http://www.mysql.com/
[14] http://www.postgresql.org/
[15] http://pl.wikipedia.org/wiki/DDoS
[16] http://developer.skype.com/Docs/ApiDoc/
[17] http://www.freedesktop.org/wiki/Software/dbus
[18] http://www.x.org/
[19] http://jabberd2.xiaoka.com/
[20] http://www.feedparser.org/

INFOBOT.PL - AN INFORMATIVE BOT

FILIP KWIATKOWSKI, MACIEJ SZEWCZYK

Abstract. The article presents Infobot - the most popular informative bot in Poland. The authors describe the way that bot works, aggregation of data and making it accessible to the users. The work concentrated on problems possible to encounter creating similar solutions and how to solve them.