Uniwersytet Mikołaja Kopernika
w Toruniu

Wydział Matematyki i Informatyki

Wydział Fizyki, Astronomii
i Informatyki Stosowanej

Tomasz Kojm

Wizualizacja dopasowywania wyrażeń regularnych w języku Perl

Praca magisterska napisana
pod kierunkiem
dra hab. Jacka Kobusa

Toruń 2005

Spis treści

Wstęp **4**

1 Wyrażenia regularne **7**
1.1 Podstawowe pojęcia lingwistyki matematycznej 7
1.2 Automaty skończone 9
1.2.1 DFA 9
1.2.2 NFA 11
1.2.3 Automaty skończone a wyrażenia regularne 12
1.2.4 NFA kontra DFA 13

2 Wyrażenia regularne języka Perl **16**
2.1 Podstawowe pojęcia 16
2.1.1 Dopasowanie wyrażenia regularnego 17
2.2 Charakterystyka wyrażeń regularnych Perla 17
2.2.1 Metaznaki 18
2.2.2 Znaki specjalne 19
2.2.3 Odwołania wsteczne 20
2.2.4 Inne rozszerzenia wyrażeń regularnych 21
2.3 Mechanizm wyrażeń regularnych 22
2.3.1 Operator dopasowania wzorca 22
2.3.2 Działanie mechanizmu dopasowania wzorca 23

3 Budowa programu **32**
3.1 Plik wykonywalny 32
3.1.1 Opcje startowe 33
3.1.2 Plik konfiguracyjny 33
3.2 Interfejs 35
3.2.1 Okno główne 35
3.2.2 Okno wyrażenia regularnego 36
3.2.3 Okno tekstu 36
3.2.4 Debug 36

3.2.5 Okno pomocy 37
3.2.6 Klawisze funkcyjne 37
3.3 Działanie 38

4 Implementacja programu **42**
4.1 Zastosowane narzędzia 42
4.1.1 Interfejs użytkownika 42
4.1.2 Dokumenty XML 42
4.1.3 Linia poleceń 43
4.2 Działanie 43
4.2.1 Parser plików XML 45
4.2.2 Okna 46
4.2.3 Kompilacja 47
4.2.4 Analiza 47
4.2.5 Wizualizacja 53
4.3 Problemy techniczne 55

Podsumowanie **57**

Spis literatury **58**

Dodatki **59**

Wstęp

Mechanizm wyrażeń regularnych to potężne narzędzie analizy danych, stanowiące podstawę zaawansowanych edytorów tekstu, narzędzi bazodanowych, języków programowania i wielu innych programów, których głównym zadaniem jest przetwarzanie informacji. Pojedyncze wyrażenia regularne pozwalają opisywać złożone zbiory napisów i tym samym zaoszczędzić czas oraz zwiększyć wydajność i skuteczność pracy, zwłaszcza z dużymi ilościami danych.

Jednym ze znamienitych przykładów oprogramowania, w którym wyrażenia regularne odgrywają kluczową rolę, jest język Perl. Wykorzystuje on je na kilka sposobów, z których najważniejszym jest proces dopasowania wzorca, polegający na porównywaniu i próbie dopasowania go do fragmentu danego tekstu. Użycie wyrażeń regularnych w charakterze wzorców pozwala na poszerzenie i większą elastyczność pola poszukiwań.

W ciągu dziesięciu lat rozwoju, mechanizm wyrażeń regularnych Perla zdążył wyewoluować do tego stopnia, że stał się jedną z większych trudności w nauce programowania w tym języku. Wykorzystanie Perla jest ściśle związane z przetwarzaniem tekstu i często jest on podstawowym narzędziem administratorów systemów komputerowych, którym pozwala na efektywne zarządzanie, analizę dzienników systemowych czy dynamiczne generowanie stron WWW. Język ten sprawdza się także doskonale w pracy z dużymi bazami danych oraz edycji i przekształcaniu tekstu. Niemal we wszystkich programach tworzonych w Perlu, główną rolę odgrywają wyrażenia regularne i nie ma wątpliwości, że to one przyczyniły się do jego olbrzymiej popularności i stanowią o jego potędze. Głębsze zrozumienie ich działania pozwala na wydajniejsze programowanie i rozwiązywanie problemów z użyciem Perla, dlatego powinno być pierwszym i bardzo istotnym krokiem w jego nauce.

Celem pracy było stworzenie narzędzia pozwalającego na obserwację działania mechanizmu wyrażeń regularnych w procesie dopasowania wzorca w Perlu. Program powinien przyspieszyć naukę wyrażeń regularnych poprzez wizualizację ich działania zarówno na odpowiednio dobranych przykładach dostarczonych wraz z programem, jak i dowolnych podanych przez użyt-

kownika. Głównymi założeniami była łatwość obsługi i zapewnienie pełnej interaktywność pomiędzy użytkownikiem a programem.

Działanie programu **REVIS** (*Regular Expression VISualiser*) opiera się na analizie danych dostarczonych przez interpreter Perla uruchomiony w trybie debugowania wyrażeń regularnych. Dostarcza on wielu informacji na temat budowy i działania mechanizmu dopasowania wzorca. Kluczową informacją są linie opisujące stan niedeterministycznego automatu skończonego dla każdego etapu dopasowania. REVIS umożliwia krokową wizualizację w obu kierunkach (następne i poprzednie dopasowanie), a także skok do wybranego stanu automatu. Na każdym etapie wizualizacji w oknie tekstu zaznaczony jest dopasowany już tekst oraz prezentowane aktualne dopasowanie, a w oknie wyrażenia regularnego odpowiedni fragment biorący w nim udział. Program używa kolorów oraz specjalnych trybów wyświetlania (np. migający tekst) dla odróżnienia różnych czynności oraz wyników dopasowania.

REVIS został wyposażony w intuicyjny interfejs kontrolowany za pomocą klawiatury komputera. Ekran główny został podzielony na cztery okna, które umożliwiają wpisanie i edycję wyrażenia regularnego oraz tekstu, podgląd działania mechanizmu i opisu aktualnej czynności. Za pomocą pliku konfiguracyjnego możliwa jest m.in. zmiana rozmiaru okien oraz dostosowanie schematu kolorów.

Całość oprogramowania została napisana w języku Perl, dzięki któremu kod analizujący dane debuggera mógł zostać zaimplementowany efektywnie oraz przejrzyście. REVIS wykorzystuje technologię XML do obsługi plików z przykładami wyrażeń regularnych i w tym celu używa popularnego modułu XML::Parser. Do stworzenia tekstowego interfejsu użytkownika wykorzystany został moduł Curses. W dobie interfejsów graficznych, interaktywne programy działające w trybie tekstowym należą do rzadkości, jednak wiele ich walorów, a zwłaszcza niskie wymagania sprzętowe, pozwalają na szerokie ich wykorzystanie, np. na terminalach tekstowych, często spotykanych na korytarzach uczelni i szkół.

Układ pracy jest następujący: w rozdziale pierwszym wprowadzone zostają podstawy matematyczne wyrażeń regularnych i automatów skończonych. Rozdział drugi poświęcono omówieniu możliwości wyrażeń regularnych języka Perl oraz szczegółowej analizie działania interpretera Perla z włączonym trybem debugowania procesu dopasowywania wzorca. W trzecim rozdziale opisana zostaje budowa oraz funkcje programu oraz przedstawione przykłady jego użycia. Rozdział czwarty zawiera szczegóły na temat implementacji programu REVIS. Praca kończy się podsumowaniem, po którym umieszczone zostały dodatki: plik konfiguracyjny oraz krótka dokumentacja w języku angielskim.

Do pracy została dołączona płyta CD-ROM, na której umieszczone zo-

stały źródła pracy w formacie LaTeX, program REVIS oraz przykłady wyrażeń regularnych. Oprogramowanie zostało udostępnione na zasadach licencji GNU General Public License[1].