Autômatos e Algoritmos de Busca de Padrões

Orientando: Rodrigo Nonamor Pereira Mariano de Souza (bolsista FAPESP)
Orientadora: Profa. Dra. Nami Kobayashi

Resumo

Apresentamos uma descrição geral do problema de busca de padrões em textos, o objeto de nossa pesquisa. Fazemos uma discussão sobre o interesse no estudo do problema, e o relacionamos com a área da Teoria dos Autômatos.

Descrição da pesquisa

Embora os dados possam ser armazenados de diversas maneiras, textos continuam sendo a principal forma de troca de informação. Por exemplo, na Literatura e na Linguística os dados são armazenados em uma vasta coleção de escritos e dicionários. Na Ciência da Computação, grande quantidade de dados são armazenados em arquivos seqüenciais. Em Biologia Molecular, moléculas biológicas podem ser aproximadas por seqüencias de nucleotídeos e aminoácidos. Dessa forma, torna-se interessante o estudo de algoritmos eficientes para a manipulação de dados textuais. Em particular, algorimos de busca em textos.

Algoritmos de busca de padrões são relevantes em três aspectos diferentes: eles são componentes básicos utilizados na implementação de comandos existentes em muitos sistemas operacionais; eles enfatizam métodos de programação que servem de paradigma em outros campos da Computação; eles desempenham papel importante na Ciência da Computação fornecendo soluções para problemas desafiadores.

Nosso projeto consiste no estudo do seguinte problema:

Dados um padrão x e um texto y, encontrar todas as ocorrências de x em y.

Um texto é meramente uma seqüência de símbolos, como este relatório ou uma cadeia de DNA. Um padrão é uma palavra ou um conjunto de palavras. Encontrar ocorrências significa encontrar posições do texto em que a palavra ou uma das palavras do conjunto apareçam.

O problema, aparentemente simples, vem sendo estudado desde a década de 70. Vários algoritmos foram propostos ao longo das décadas de 70, 80 e 90, e o problema ainda é objeto de pesquisa. Alguns desses algoritmos encontram-se implementados em aplicações de busca em sistemas operacionais, editores de texto, *software* de manipulação de imagens e de cadeias biológicas.

Várias das soluções propostas são sustentadas em resultados da Teoria dos Autômatos. Nessa área são estudadas propriedades combinatórias de palavras e formalismos para a manipulação de palavras. Em particular, há o formalismo de *autômato finito*, que serve de base para vários algoritmos de busca de padrão. Esse formalismo originou-se de estudos biológicos de redes de neurônios e circuitos de chaveamento. Atualmente, é utilizado, por exemplo, no desenvolvimento de compiladores, editores de texto e sistemas de manipulação de arquivos.

Um autômato finito pode ser visto como uma máquina que pode assumir vários estados de um conjunto finito. A única informação que essa máquina pode armazenar é o estado em que se encontra. Um autômato funciona recebendo uma palavra e fazendo

a leitura seqüencial dessa palavra. A cada símbolo lido, ocorre uma mudança de estado. Após a leitura da palavra o autômato indica se ela foi aceita, caso ela verifique uma determinada propriedade.

Há diversas maneiras de utilizar um autômato em busca de padrões. Há algoritmos que constroem um autômato sobre o padrão e posteriormente a simulam esse autômato no texto. Durante essa simulação, os estados assumidos pelo autômato indicam a situação da busca, representando quais partes do padrão foram encontradas no trecho do texto já lido. Uma outra abordagem é considerar o texto fixo e, com um pré-processamento sobre o texto, construir um índice através do qual é possível fazer uma série de buscas de maneira eficiente. Um dos métodos de representar esse índice é com autômatos finitos.

Em nossa pesquisa nos concentramos nas aplicações de autômatos finitos em problemas de busca de padrão. Estudamos alguns resultados da Teoria dos Autômatos e descrevemos alguns algoritmos de busca que fazem uso de um autômato, incluindo uma descrição de suas complexidades. Alguns tópicos desse estudo são: Algoritmo de Knuth, Morris e Pratt, Algoritmo de Simon, Algoritmo de Boyer e Moore, Árvore dos Sufixos e Autômato dos Sufixos. Estudaremos ainda busca de padrões com aproximação, quando não se deseja encontrar ocorrências exatas do padrão, e busca em textos bidimensionais. Faremos a implementação e testes de alguns algoritmos. As implementações e maiores informações poderão ser encontradas em http://www.linux.ime.usp.br/~rsouza/.

Referências

1.A. V. Aho. *Algorithms for Finding Patterns in Strings*. Handbook of Theoretical Computer Science, vol. A, cap. 5, pag. 255-300, Elsevier, 1990.

2.A. V. Aho, M. J. Corasick. *Efficient String Matching: an Aid to Bibliographic Search*. Comm. ACM, 18:333-340, 1975.

3.R. S. Boyer, J. S. Moore. A Fast String Searching Algorithm. Comm. ACM, 25:762-772, 1977.

4.M. Crochemore, R. Ryter. *Text Algorithms*. Oxford University Press, 1994.

5.M. Crochemore, C. Hancart. *Pattern Matching in Strings*. Algorithms and Theory of Computation Handbook cap. 11, CRC Press, Boca Raton, 1998.

6.M. Crochemore, C. Hancart. *Automata for Matching Patterns*. Handbook of Formal Languages, vol. 2 em Linear Modeling, pag. 399-462, Springer-Verlag, 1997.

7.D.E. Knuth, J. H. Morris Jr., V. R. Pratt. *Fast Pattern Matching in Strings*. SIAM J. Comp., 6:323-350, 1977.

8.P. F. B. Menezes. *Linguagens Formais e Autômatos*. Ed. Sagra Luzzatto, 1997.

9.K. Thompson. Regular Expressions Search Algorithms. Comm. ACM 11:419-422, 1968.

10.I. Simon. *String Matching Algorithms and Automata*. Results and Trends in Theoretical Computer Science, n.o 812 em LNCS, pag. 386-395, Springer-Verlag, 1994.