

Estruturas de Dados e Algoritmos para Busca de Padrões

Proposta de Trabalho

Marco Alves de Alcantara, Cristina Gomes Fernandes

6 de Maio de 2022

1 Descrição

O nosso objetivo nesse TCC é o estudo de busca de padrões em texto de uma maneira computacionalmente eficiente, isso é, dado um texto qualquer, encontrar todas as ocorrências de uma palavra (ou várias palavras sucessivamente), bem como contar esse número de ocorrências dentro daquele texto.

Uma possível utilidade disso para seres humanos é quando, por exemplo, tem-se um documento de texto ou uma página web genérica, e bastante longa, e deseja-se saber se uma palavra ocorre no documento, ou então procurar onde exatamente ela está, sem que seja necessário buscar manualmente.

Também iremos implementar algoritmos e as estruturas de dados que possibilitam uma solução eficiente para esse problema, e escreveremos sobre esse problema, sobre os algoritmos e as implementações desenvolvidas.

2 Estruturas de dados para busca de padrão

Uma das estruturas de dados mais importantes na solução desse problema é o chamado vetor de sufixos. Trata-se de um vetor de inteiros, que indexa o texto onde deseja-se realizar as buscas. O vetor é ordenado alfabeticamente com relação aos sufixos do texto iniciados a partir daquele índice [Gus97, pp. 149–151].

Para utilizá-lo eficientemente, também é necessário o vetor LCP (de *Longest Common Prefix*), que guarda o comprimento do maior prefixo entre dois sufixos consecutivos do vetor de sufixos. Esse vetor, embora não seja necessário para a busca propriamente dita, serve para torná-la mais eficiente, pois é possível pular algumas comparações entre caracteres da palavra e do texto [Gus97, p. 152].

Há, também uma generalização do vetor LCP, os vetores LCP estendidos (RLCP e LLCP). Eles são úteis pois, ao contrário do vetor LCP tradicional, que guarda apenas o comprimento dos prefixos de sufixos consecutivos do texto, o RLCP e LLCP guardam o comprimento entre um par de posições que pode surgir durante uma busca binária pela palavra no vetor de sufixos, tornando o algoritmo ainda mais eficiente [Gus97, pp. 152–155]. Todos esses vetores podem ser construídos em tempo linear com relação ao comprimento do texto [Kas+01], [Gus97, p. 152], [KS06].

Para esse TCC, todas as funções de busca usando esses vetores já foram implementadas, bem como a construção linear de todos, com exceção do vetor de sufixos, para o qual ainda usamos o algoritmo trivial de construção (de tempo quadrático). Um dos próximos passos é justamente implementar a construção do vetor de sufixos em tempo linear.

A outra estrutura de dados muito usada nesse contexto é a árvore de sufixos, que também pode ser usada para realizar as buscas num texto [Gus97, pp. 90–93]. Ela pode ser construída com um algoritmo iterativo através de árvores cartesianas a partir do vetor LCP, em tempo linear [Dem21]. Sua construção, bem como suas funções de busca já foram implementadas nesse TCC. Pretendemos implementar sua construção em tempo linear a partir dos vetores de sufixos e LCP, bem como sua construção linear direta sem esses vetores.

3 Cronograma aproximado

A maioria dos algoritmos e as funções de busca já foram estudados, e a maioria deles já foi implementada na disciplina MAC0385 (Estruturas de Dados Avançadas), ministrada no segundo semestre de 2021. Iremos revisar, e possivelmente aprimorar as implementações que já foram feitas. Além disso, falta implementar a construção em tempo linear do vetor de sufixos de um texto, e também estudar e implementar algoritmos lineares para construção da árvore de sufixos sem o vetor LCP (como o Algoritmo de Ukkonen) [Gus97, pp. 94–107].

- Maio: revisão e implementação da construção linear do vetor de sufixos;
- Maio e Junho: revisão dos algoritmos implementados até o momento e escrita e revisão da parte da monografia sobre o vetor de sufixos;
- Julho: estudo e implementação dos algoritmos de conversão entre árvores de sufixos e vetor de sufixos;
- Agosto: escrita e revisão da parte da monografia sobre esses algoritmos de conversão; estudo do algoritmo de Ukkonen, de construção linear da árvore de sufixos;
- Setembro: implementação do algoritmo de Ukkonen;
- Setembro e Outubro: escrita e revisão da parte da monografia sobre o algoritmo de Ukkonen;

- Outubro: estudo de alguma aplicação extra de árvores de sufixos e/ou vetor de sufixos;
- Novembro: escrita e revisão da parte da monografia sobre esta parte extra; preparação do pôster e da apresentação do TCC;
- Novembro e Dezembro: finalização da escrita da monografia.

Referências

- [Gus97] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences*. 1997, pp. 94–107.
- [Kas+01] Toru Kasai et al. “[Linear-Time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications](#)”. Em: (2001).
- [KS06] Juha Kärkkäinen e Peter Sanders. “[Simple Linear Work Suffix Array Construction](#)”. Em: (2006).
- [Dem21] Erik Demaine. *Advanced Data Structures (Spring ’21) Lecture Notes*. [Lecture 15](#) and [Lecture 16](#). 2021.