5.4 Pesquisa Digital

A pesquisa digital é baseada na representação das chaves como uma sequência de caracteres ou de dígitos. Grosso modo, o método de pesquisa digital é realizado da mesma forma que uma pesquisa em dicionários que possuem aqueles "índices de dedo". Com a primeira letra da palavra são determinadas todas as páginas que contêm as palavras iniciadas por aquela letra.

Os métodos de pesquisa digital são particularmente vantajosos quando as chaves são grandes e de **tamanho variável**. No problema de casamento de cadeias, trabalha-se com **chaves semi-infinitas**[4], isto é, sem limitação explícita quanto ao tamanho. Um aspecto interessante quanto aos métodos de pesquisa digital é a possibilidade de localizar todas as ocorrências de determinada cadeia em um texto, com tempo de resposta logarítmico em relação ao tamanho do texto.

5.4.1 Trie

Uma trie é uma árvore M-ária cujos nós são vetores de M componentes com campos correspondentes aos dígitos ou caracteres que formam as chaves. Cada nó no nível i representa o conjunto de todas as chaves que começam com a mesma sequência de i dígitos ou caracteres. Esse nó especifica uma ramificação com M caminhos dependendo do $(i+1)$-ésimo dígito ou caractere de uma chave. Considerando as chaves como sequência de *bits* (isto é, $M = 2$), o algoritmo de pesquisa digital é semelhante ao de pesquisa em árvore, exceto pelo fato de que, em vez de se caminhar na árvore de acordo com o resultado de comparação entre chaves, se caminha de acordo com os *bits* de chave. A Figura 5.9 mostra uma trie construída a partir das seguintes chaves de 6 *bits*:

B = 010010
C = 010011
H = 011000
J = 100001
M = 101000

Para construir uma trie, faz-se uma pesquisa na árvore com a chave a ser inserida. Se o nó externo em que a pesquisa terminar for vazio, cria-se um nó externo nesse ponto contendo a nova chave, como ilustra a inserção da chave W = 110110 na Figura 5.10. Se o nó externo contiver uma chave, cria-se um ou mais nós internos cujos descendentes conterão a chave já existente e a nova chave. A

[4]Uma chave semi-infinita é uma sequência de caracteres em que somente a sua extremidade inicial é definida. Logo, cada posição no texto representa uma chave semi-infinita, constituída pela sequência que inicia naquela posição e se estende à direita tanto quanto for necessário ou até o final do texto. Por exemplo, um banco de dados constituído de n palavras (as posições de interesse nesse caso são os endereços de início das palavras) possui n chaves semi-infinitas.

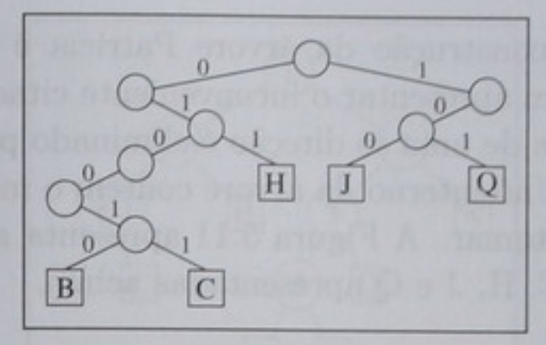

Figura 5.9 *Trie binária.*

Figura 5.10 ilustra a inserção da chave K = 100010 que envolve repor J por um novo nó interno cuja subárvore esquerda é outro novo nó interno, cujos filhos são J e K, porque estas chaves possuem os mesmos *bits* até a quinta posição.

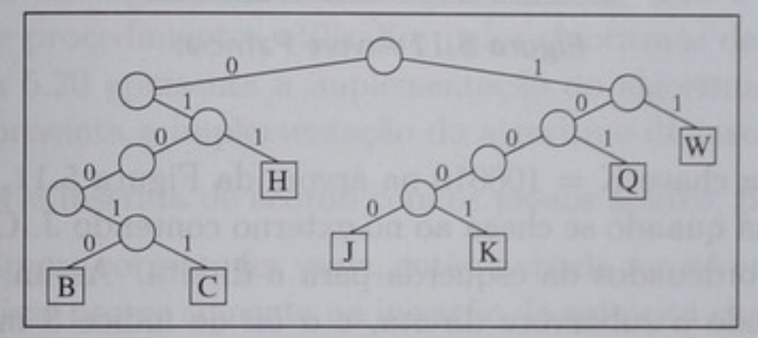

Figura 5.10 *Inserção das chaves W e K.*

O formato das tries, diferentemente das árvores binárias comuns, não depende da ordem em que as chaves são inseridas, e sim da estrutura das chaves por meio da distribuição de seus *bits*. Uma grande desvantagem das tries é a formação de caminhos de uma só direção para chaves com um grande número de *bits* em comum. Por exemplo, se duas chaves diferirem somente no último bit, elas formarão um caminho cujo comprimento é igual ao tamanho delas, não importando quantas chaves existem na árvore. Veja o caminho gerado pelas chaves B e C na Figura 5.10.

5.4.2 Patricia

PATRICIA é a abreviatura de Practical Algorithm To Retrieve Information Coded In Alphanumeric (Algoritmo Prático para Recuperar Informação Codificada em Alfanumérico). Esse algoritmo foi originalmente criado por Morrison (1968) em um trabalho aplicado à recuperação de informação em arquivos de grande porte. Knuth (1973) deu um novo tratamento ao algoritmo, reapresentando-o de forma mais clara como um caso particular de pesquisa digital, essencialmente um caso de árvore trie binária. Sedgewick (1988) apresentou novos algoritmos de pesquisa e de inserção baseados nos algoritmos propostos por Knuth (1973). Gonnet e Baeza-Yates (1991) também propuseram outros algoritmos.