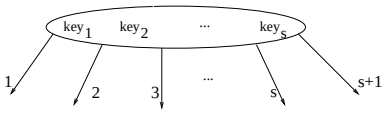
CS/ENGRI 172, Fall 2002
10/9/02: Lecture Eighteen Handout

Topics: Indexing using B-trees; introduction to the vector-space model.

B-trees

B-trees (*balanced multiway search trees*) help make the term lookup process more efficient. Conceptually, you can think of a B-tree as sitting "on top" of an index, with each leaf corresponding to the information for some single term in the index. (Strictly speaking, when the leaves are data items we have a B^+*-tree* rather than a B-tree, but we won't make this distinction.)

Every B-tree has some *order* (or *minimization factor*) t such that except for the root, each internal node contains between t and $2t$ *keys* in sorted order, for example,



where s is some number such that $t \leq s \leq 2t$. (The root itself is an exception; it can have between 1 and $2t$ keys.)

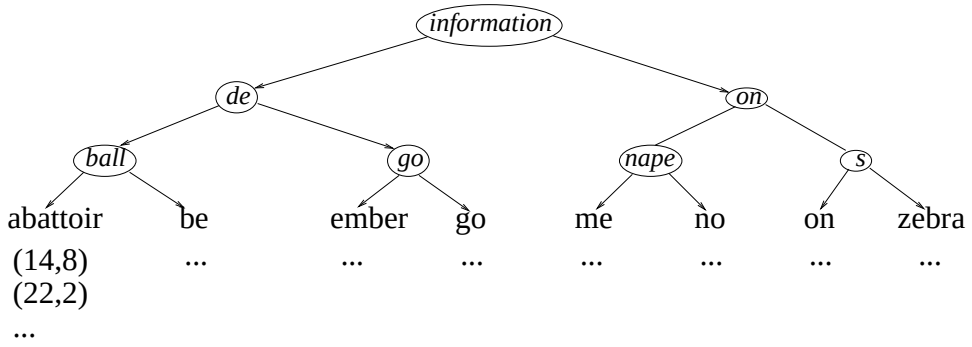
The keys give information about the leaves of the $s+1$ *subtrees*. The i^{th} child "covers" terms w such that $\text{key}_{i-1} \preceq w \prec \text{key}_i$. The exceptions are the 1st child, which "covers" terms w such that $w \prec \text{key}_1$, and the $s+1$th child, which "covers" terms w such that $\text{key}_s \preceq w$.

Finally, we require that every leaf of the B-tree have the same depth.

A crucial fact about B-trees for our indices is that the depth of such B-trees can be at most roughly $\log_t(m)$.

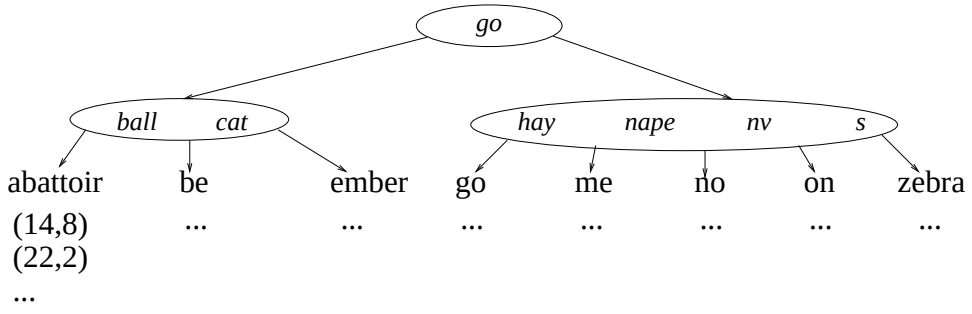
Example B-trees

Here is a B-tree of order 1.



(OVER)

And here is a B-tree of order 2 for the same index.



Vector length normalization

Recall that for any vector $\overrightarrow{x} = (x_1, x_2, \ldots, x_n)$, the vector length of x is $\sqrt{\overrightarrow{x} \cdot \overrightarrow{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$. A handy fact is that we can *normalize* any vector of non-zero length to get a new vector that points in the same direction, but has unit length. We do this simply by dividing each component of the old vector by the old vector's length (call this L):

$$\begin{aligned}
\text{length}\left((x_1/L, x_2/L, \ldots, x_n/L)\right) &= \sqrt{x_1^2/L^2 + x_2^2/L^2 + \cdots + x_n^2/L^2} \\
&= \frac{1}{L}\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \\
&= \frac{1}{L}\text{length}\left(\overrightarrow{x}\right) \\
&= \frac{1}{L}(L) = 1.
\end{aligned}$$