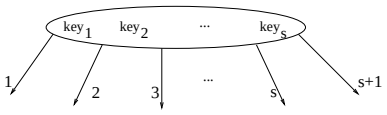
Computation, Information, and Intelligence (COMS/ENGRI/INFO/COGST 172), Fall 2005
9/28/05: **Lecture 15 aid – B-trees; the vector-space model**

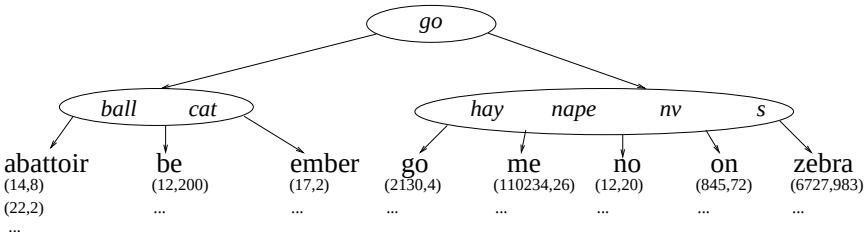
Agenda: Finish B-trees; start discussion of the vector-space model.
Announcements: Due to the **in-class prelim on October 7th**, we will not be handing out a new homework assignment before then so as to allow you to concentrate on preparing for the exam. We will also be temporarily altering the office-hours schedule so that there are consultation times available closer to the exam date. More information is forthcoming, but given that Homework Two was due today, it seems that a little breather before talking about the next class "event" is in order.

I. Reminder: the main idea behind B-trees *Make sure to look at the previous lecture's aid for all the technical details, especially the definition of a B-tree's order.* Internal (i.e., non-leaf) B-tree nodes look like this:



The keys must be distinct and in sorted order. The i^{th} child's subtree "covers" terms w such that $\text{key}_{i-1} \preceq w \prec \text{key}_i$. The exceptions are the first child, whose subtree "covers" terms w such that $w \prec \text{key}_1$, and the $s+1$th child, whose subtree "covers" terms w such that $\text{key}_s \preceq w$.

II. An order-2 B-tree



(OVER)

III. Normalized term-frequency vectors

Define $\text{freq}_{i,j}$ (term-document frequency) as the number of times term w_i occurs in document d_j. We then set the document vector $\overrightarrow{d}_j$ for document d_j as follows:

$$\overrightarrow{d}_j = \left(\frac{\text{freq}_{1,j}}{N_j}, \frac{\text{freq}_{2,j}}{N_j}, \ldots, \frac{\text{freq}_{m,j}}{N_j} \right)$$

where $N_j = \sqrt{\sum_{i=1}^{m} \text{freq}_{i,j}^2}$ is the length-normalization factor.

IV. Example corpus and query

Vocabulary: w_1: cats; w_2: dogs; w_3: news

d_4: "cats news"
d_5: "cats news cats news"
d_6: "cats dogs news news dogs"

q: "cats dogs"