Topics: more on B-trees; retrieval based on the vector-space model, allowing the incorporation of more information regarding term frequencies.

Follow-ups to last time: If you query Google for “how tall is mount everest”, you receive a specialized response:

Mount Everest — Elevation: 8,850 M (29,035 feet) Ranked 1st
According to http://copernicus.subdomain.de/Mount_Everest - More sources }}

But what if you query Google for “how tall is greek peak”?

Nothing doing.

I. Example B-tree Remember that the index is stored (in sorted order, left to right) in the leaves.

\[
\begin{array}{c}
\text{go} \\
\text{ball} & \text{cat} \\
\text{abattoir} & \text{be} & \text{ember} \\
(14,8) & \text{...} & \text{...} \\
(22,2) & \text{...} & \text{...} \\
\text{hay} & \text{nape} & \text{nv} & \text{s} \\
\text{go} & \text{me} & \text{no} & \text{on} & \text{zebra} \\
\text{...} & \text{...} & \text{...} & \text{...} & \text{...} \\
\end{array}
\]

This tree is of order \( t = 2 \) because the root contains between 1 and \( 2 \cdot 2 \) keys inclusive, and the other internal nodes contain between 2 and \( 2 \cdot 2 \) (inclusive) keys. See previous lecture aid for definitions.

II. B-tree search (See that you understand from the definitions given on lecture aid from last time why this procedure works.) Suppose we want to know which documents contain the term \( w \). Begin at the left-hand side of the root. Move rightwards among the (sorted) keys until you either first hit a key \( k \neq w \) such that \( k \) alphabetically follows \( w \) or you hit the right end of the node; then, follow the child that you are “at”. Repeat until you hit a leaf.

III. Example data Let the vocabulary \( W \) be \( w_1: \text{cat}; w_2: \text{dog}; w_3: \text{news} \).

A three-document corpus: \( d: \text{“cat news”} \) \( d': \text{“news cat news cat news”} \) \( d'': \text{“cat dog news dog news”} \)

Query \( q: \text{“cat dog”} \)

IV. Normalized term-frequency vectors Let the terms be \( w_1, w_2, \ldots, w_m \). Define the term-document frequency \( \text{freq}(w_i \in d) \) as the number of times \( w_i \) occurs in \( d \). We then set the document vector \( \overrightarrow{d} \) for document \( d \) as follows:

\[
\overrightarrow{d} = \left( \frac{\text{freq}(w_1 \in d)}{N(d)}, \frac{\text{freq}(w_2 \in d)}{N(d)}, \ldots, \frac{\text{freq}(w_m \in d)}{N(d)} \right)
\]

where \( N(d) = \sqrt{\sum_{i=1}^{m} \text{freq}(w_i \in d)^2} \) is the length-normalization factor.

\[1\] 2100 feet.