索引的加速

信息论

信息量

信息量用于衡量信息的多少
信息量 = 事件与预期的差别程度(surprisal)

信息熵

信息熵(Entropy)=信息量的期望值

> Wikipedia: Information entropy is a concept from information theory. It tells how much information there is in an event. In general, the more uncertain or random the event is, the more information it will contain. The concept of information entropy was created by mathematician Claude Shannon.

$$\mathrm{H}(X) = \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i),$$

信息熵是对不确定性的衡量。我们对X的取值越不确定，H(X)的值越大

If one of the events is more probable than others, observation of that event is less informative. Conversely, rarer events provide more information when observed. Entropy is zero when one outcome is certain.

Generally, entropy refers to disorder or uncertainty.

编码

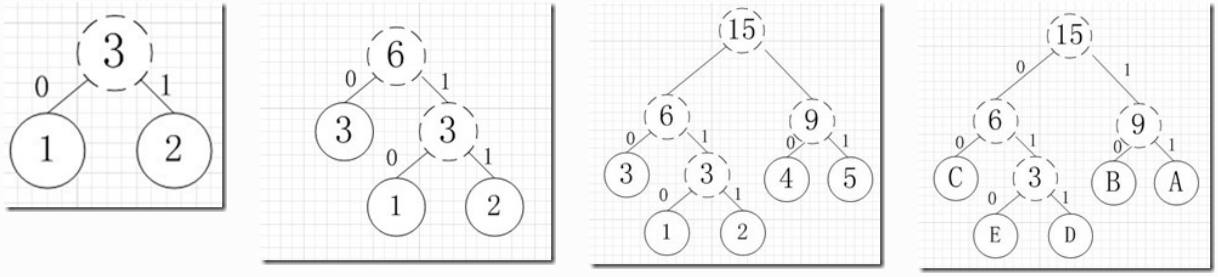
哈夫曼编码

> Wikipedia: In computer science and information theory, a Huffman code is a particular type of optimal prefix code that is commonly used for lossless data compression.

哈夫曼编码的例子

> 假如我有A,B,C,D,E五个字符，出现的频率（即权值）分别为5,4,3,2,1
>
> Step1: 取两个最小权值作为左右子树构造一个新树
>
> Step2: 再把新生成的权值为3的结点放到剩下的集合中，所以集合变成{5,4,3,3},取最小的两个权值构成新树
>
> Step3: 同上处理
>
> Step4: 其中各个权值替换对应的字符即为下图
>
> 所以各字符对应的编码为：A:11 , B:10 , C:00 , D:011 , E:010

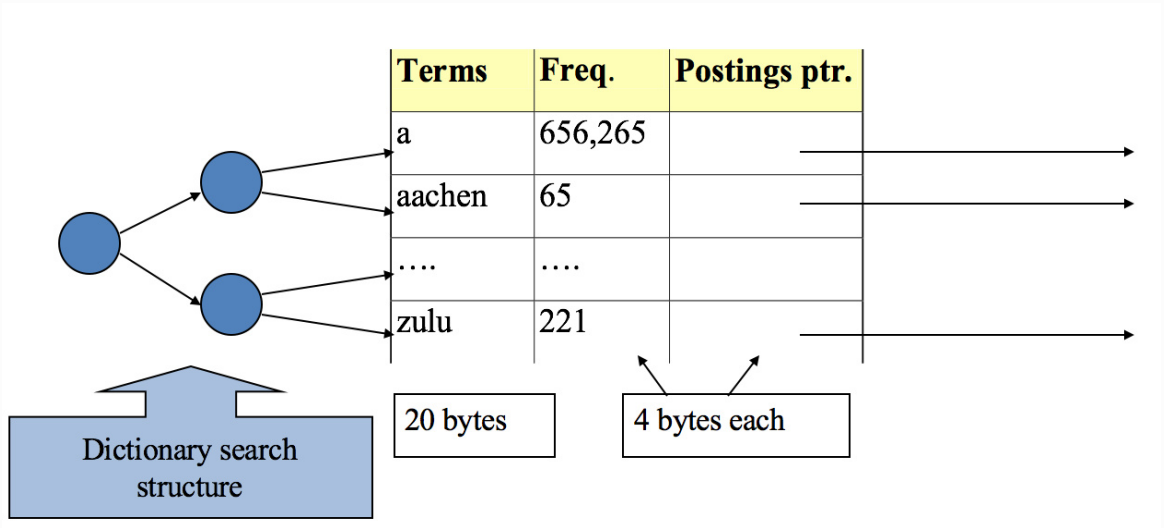
Step1 **Step2** **Step3** **Step4**



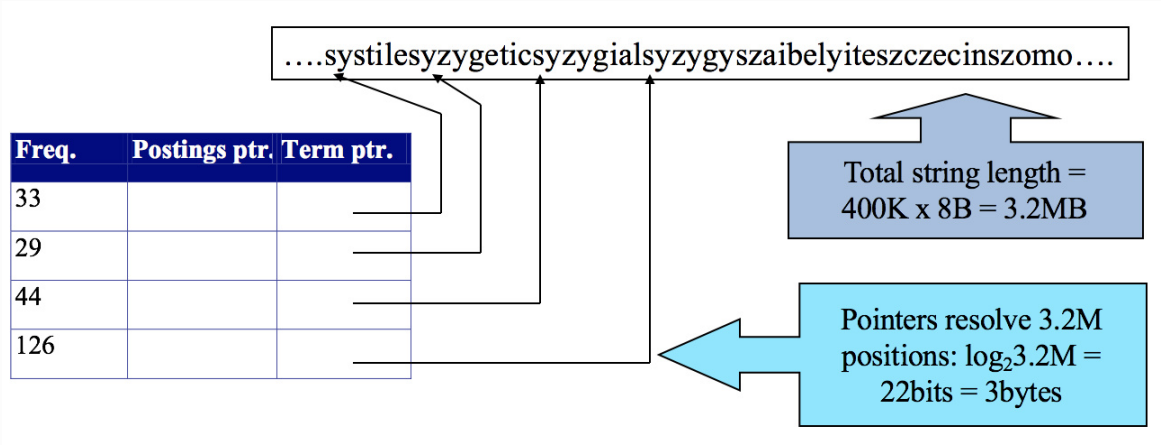
从理论上讲，对任何一类数据，都存在某种编码方式，其编码长度接近于该类数据的信息熵。

压缩方案

- dictionary, fixed-width

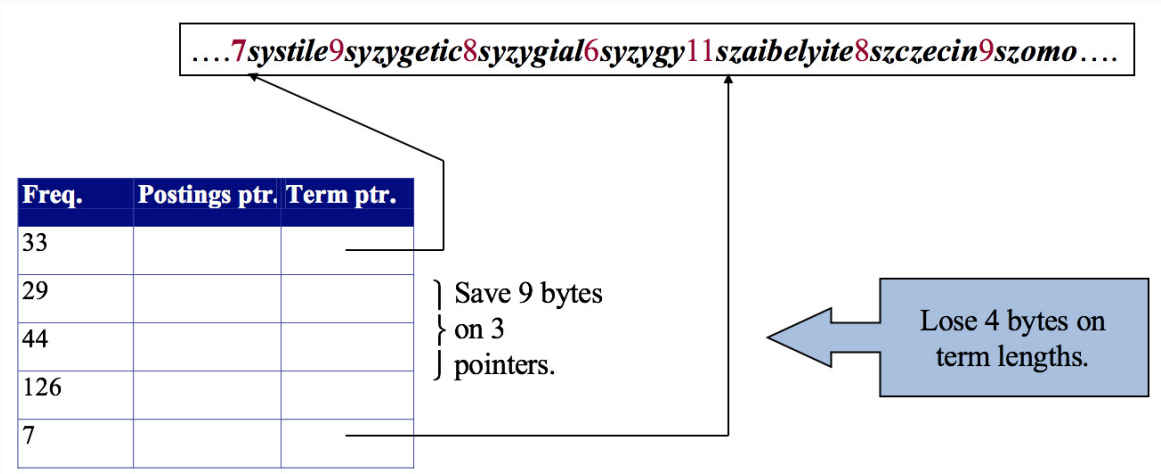


- Dictionary As a String : 去除无效空间



- Blocking : 将k(k=4)个单词合并成一个Block

> 3abc4sdef6wsdsas5dcsxsf2df4idju5udjif (Number represents the length of word)



- With blocking & front coding

压缩可能带来的问题

- 压缩可以减少对存储介质的访问，从而加速查询。但压缩的代价是计算量的增加，它也可能拖延查询。

Posting List 的压缩

1. Delta Encoding (d-gap) : 用相邻文档号的差值代替文档号

> 12 , 16 , 30 , 50 →12 , 4 , 14 , 20 (然后用二进制编码)
>
> 文档号的gap值分布非常不均，极易压缩。
>
> 短的gap频繁，长的gap稀少

2. unary codes

> Unary coding, sometimes called thermometer code, is an entropy encoding that represents a natural number, **n, with n ones followed by a zero** (if natural number is understood as non-negative integer) or with n – 1 ones followed by a zero (if natural number is understood as strictly positive integer).