WEBinfo-HW-1

PB19030888 张舒恒

1.1

OR运算在最坏情况下的运算结果长度为O(x+y)，则各词项的运算结果长度为(tangerine OR trees)：363465，(marmalade OR skies)：379571，（kaleidoscope OR eyes）：300321，AND运算应该优先选择长度短的，则处理次序为 (kaleidoscope OR eyes）AND (tangerine OR trees) AND (marmalade OR skies)

1.2

(1)跳表指针实际跳转1次，即表一的24小于表二的89且跳转位置75也小于89发生跳转

(2)倒排记录之间的比较次数是19次，即3==3,5==5,9<89,15<89,24<89,75<89,75<89,92>89,81<89,84<89,89==89,92<95,115>95,96>95,96<97,97==97,100>99,100==100,115>101

(2)倒排记录之间的比较次数是19次，即3==3,5==5,9<89,15<89,24<89,39<89,60<89,68<89,75<89,81<89,84<89,89==89,92<95,96>95,96<97,97==97,100>99,100==100,115>101

1.3

按照最后一个字节延续位1其余字节延续位0对ID的间距编码：

ID	ID的间距	编码	省略高位0后的编码
777	777	00000110 0001001	110 0001001
17743	16966	00000001 0000100 1000110	1 0000100 1000110
294068	276325	00010000 1101110 1100101	10000 1101110 1100101
31251336	30957268	00001110 1100001 0111101 1010100	1110 1100001 0111101 1010100

1.4

记状态i生成观测o_j的概率为$b_i(o_j)$，初始状态分布概率$\overrightarrow{\pi}=(\pi_1,\pi_2,\pi_3)=(0.2,0.4,0.4)$，观测序列$\overrightarrow{O}=(o_1,o_2,o_3)=(\text{晴天},\text{雨天},\text{晴天})$，最优路径$I^*=(i_1^*,i_2^*,i_3^*)$，时刻t状态为i观测序列为$(o_1,o_2,o_3)$的最大概率$\delta_t(i)$，时刻t概率最大路径的前一个时刻状态为$\psi_t(i)$。初始化$\delta_1(i)=\pi_i b_i(o_1)$，$\psi_1(i)=0$，$i=1,2,3$，递推$t=2,3$，$\delta_t(i)=max_{1\le j\le 3}[\delta_{t-1}(j)a_{ji}]b_i(o_t)$，$\psi_t(i)=argmax_{1\le j\le 3}[\delta_{t-1}(j)a_{ji}]$，$i=1,2,3$，列表如下：

$\delta_t(i)$	i=1	i=2	i=3
t=1	0.1	0.028	0.00756
t=2	0.16	0.0504	0.01008
t=3	0.28	0.042	0.0147

$\psi_t(i)$	i=1	i=2	i=3
t=1	0	3	2
t=2	0	3	2
t=3	0	3	3

最优路径终点$i_3^* = argmax_i[\delta_3(i)] = 3$，回溯$i_2^* = \psi_3(i_3^*) = 3$，$i_1^* = \psi_2(i_2^*) = 3$，则该名商人最有可能的旅行轨迹是$(3, 3, 3)$。

2.1

可以使用环形一致性hash算法，将url和网站IP地址分配到多个由不同机器组织的hash键值空间，在节点数量动态变更的情况下只需要遵循顺时针存储规则进行相邻节点的资源转移。

2.2

在查询词项分布密集处适当增长跳表指针步长，在查询词项分布稀疏处适当缩短跳表指针步长。

2.3

在已有停用词表后如果搜索到非停用词则使用位置索引进行记录，否则不进行记录。潜在问题如一些停用词在特定语境下有特殊的语义，这使得一些短语中的单词会被识别为停用词，导致无法只从单词的位置信息推断出短语。解决方案如适当用统一的记号代替停用词表中所有停用词，便可以大致推断停用词位置信息。

2.4

改进后的设计方案为先计算各个词汇在文档中的频率，设定频率阈值，从左至右/从右至左尽可能匹配最长的词，直到该词项的文档频率低于频率阈值。在此种方案下，反向最大匹配分词的效果仍优于正向最大匹配分词，因为这是自然语言左歧义频率高于右歧义频率的特性所决定的，比如“ 使用户满意 ”一词中的“ 使用户 ”正确分词为“ 使/用户 ”，同时有左歧义分词“ 使用/户 ”，而考虑词汇频率的分词方法不能改变左歧义和右歧义频率不均带来的RMM优于FMM的结果。

2.5

对于自然语言如中文和英文来说，不同词语拥有相同前缀是较为普遍的，如concatenate，concave，concavity等等，所以针对Trie树稀疏度高空间利用率低的问题，可以使用双数组Trie树，它是Trie结构的压缩形式，仅用两个线性数组来表示Trie树。由于在双数组所有键中包含的字符之间的联系都是通过简单的数学加法运算表示，所以能明显提高查询效率。