NLP算法入门系列：隐含马尔可夫链(HMM)模型的简单介绍

原创 IT可达鸭 IT可达鸭 5月6日

文/IT可达鸭

图/IT可达鸭、网络

- **前言**

随着大规模语料库的建立，以及统计学、机器学习方法的研究和发展，基于统计的中文分词算法逐渐成为主流。

- **基于统计分词的详解**

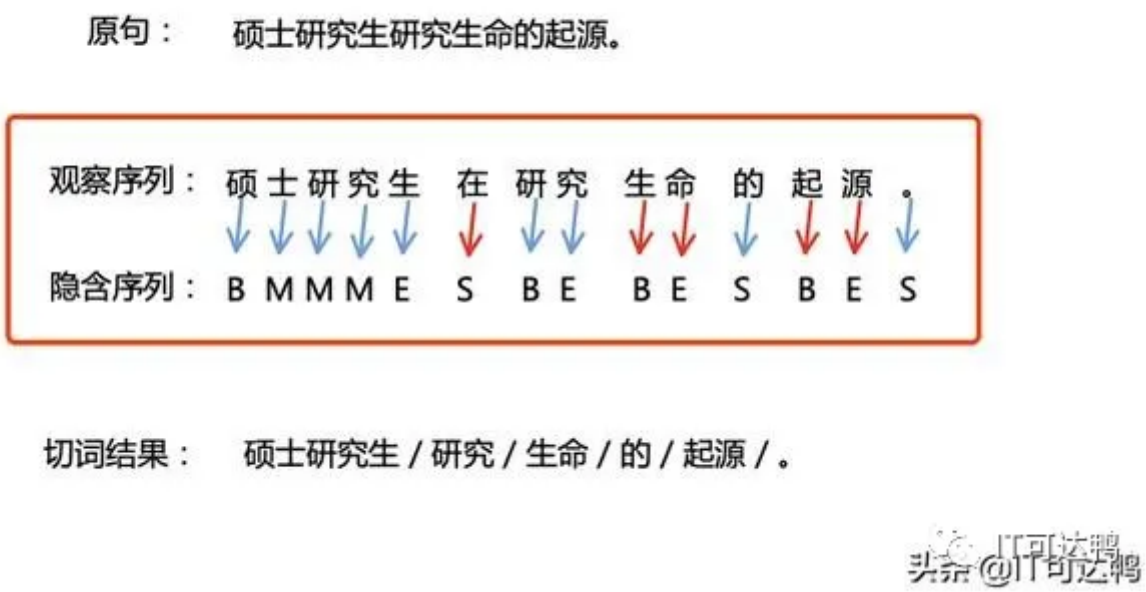
主要思想：**把每n(n>=1)个相邻的字（可重叠）看作是一个待识别的词，如果待识别的词在不同文本中出现的次数越多，就说明这待识别的词很可能就是一个词。**

因此，我们可以利用字与字相邻出现的频率来反应组成词的可靠度，统计语料中相邻共现的各个字的组合的频率，当频率高于某一个临界值的时候，便可以认为该字的组合可能是一个词。

- **基于统计的分词算法: HMM**

隐含马尔可夫模型（HMM）是将分词作为字在句子中的序列标注任务来实现分词的。其基本思路是：**每个字在构造一个特定的词语时都占据着一个确定的词位，现设定每个字最多只有四种构词位置：即B（Begin 词首）、M（Middle 词中）、E（End 词尾）、S（single 单独成词）。**

举个例子：



在HMM之前，必须得有三个假设：

假设1: 有限历史性假设，采用二元模式，每个字只与上一个字和下一个字有关联;

假设2: 观测独立性假设，输出仅与当前状态有关;

假设3: 齐次性假设;

另外HMM的标注必须满足：只有出现BE、BME、BMME、BM...ME（中间M的个数大于等于0）、S 这几种情况。

简单的理解，HMM就是通过观察序列，求解隐含序列的一个过程。

即，最大化 P(字序列|标签序列)

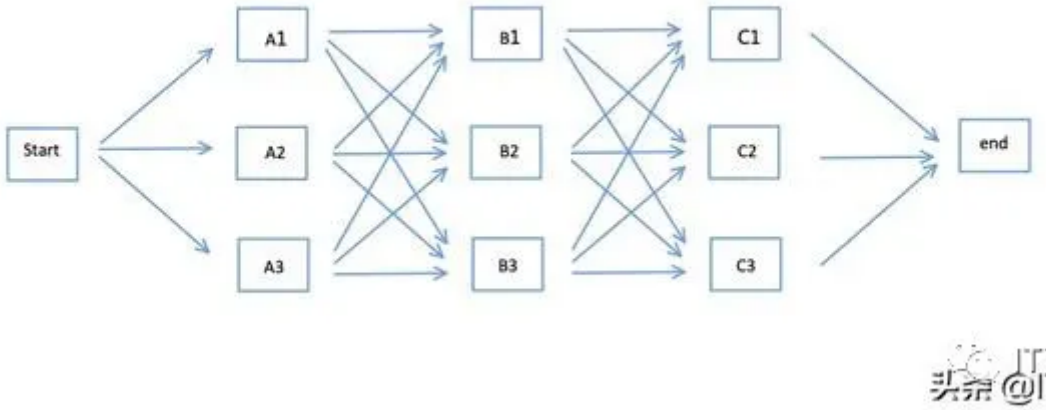
$$P(\text{字序列}|\text{标签序列}) \sim P(\text{字1}|\text{标签1}) * P(\text{标签2}|\text{标签1}) * P(\text{字2}|\text{标签2}) * P(\text{标签3}|\text{标签2}) *** P(\text{标签n} | \text{标签n}-1)P(\text{字n} | \text{标签n})$$

在求解HMM的过程中，需要维护三个矩阵：**1. 初始概率分布；2.状态转移矩阵A（标签->标签的概率）；3. 观察概率分布B（标签->观察变量的概率）**

由假设1可以知，如果最终的最优路径经过某个o(i)节点，那么从初始节点到o(i-1)点的路径必然也是一个最优路径，因为每个节点o(i)只会影响前后两个节点的标签概率。所以最大化 P(字序列|标签序列）可以通过Viterbi 算法来解决。

- **基于统计的分词算法: Viterbi**

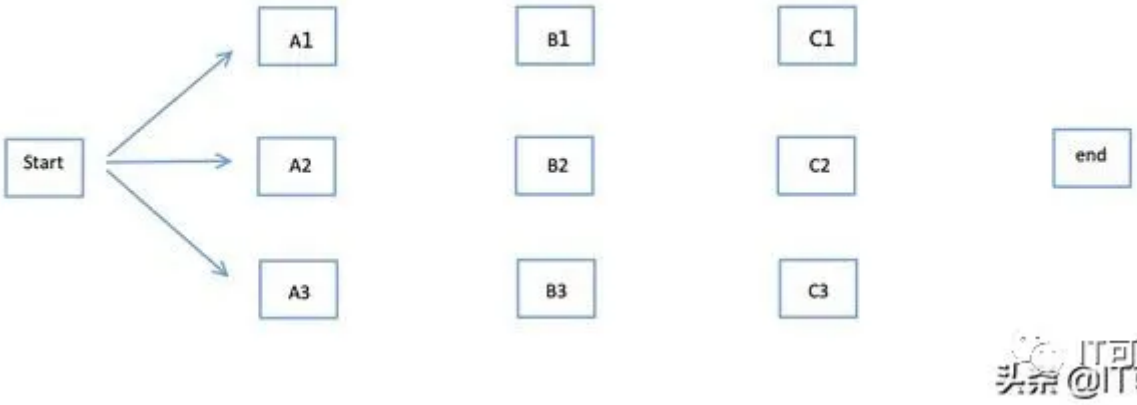
在HMM中，求解模型最常用的方法是Viterbi算法。它是一种动态规划方法，核心思想如下图所示，为了方便演示，将HMM中的BMES标注用1、2、3进行代替演示，字符用A、B、C代替演示。



基于假设1，每一列的节点只能和相邻列的节点相连，不能跨列相连，节点之间有着不同的距离。

为了找出start到end之间的最短路径，我们先从Start开始从左到右一列一列地看。

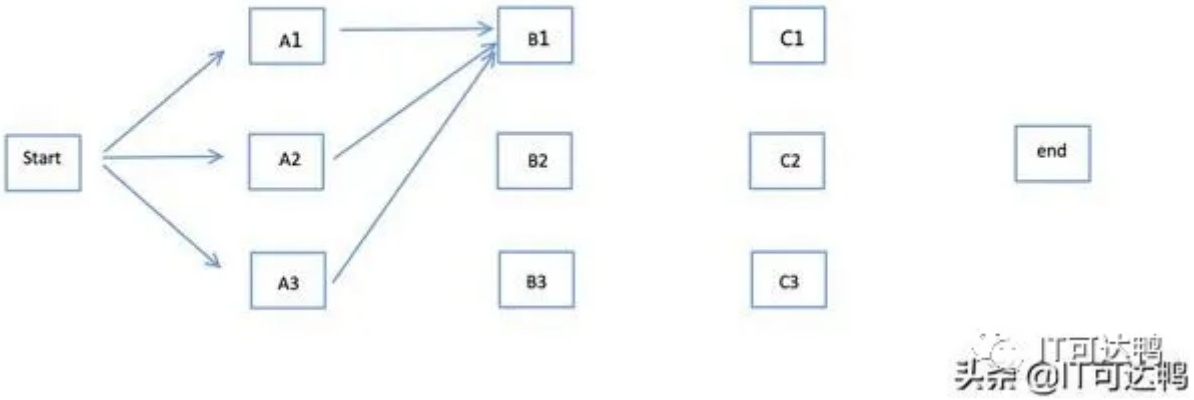
首先起点是Start，从Start到A列的路径有三种可能：Start-A1、Start-A2、Start-A3，如下图：



这时，不能武断地说Start-A1、Start-A2、Start-A3中哪一段必定是全局最短路径中的一部分，目前为止，任何一段都有可能是全局最短路径的备选。

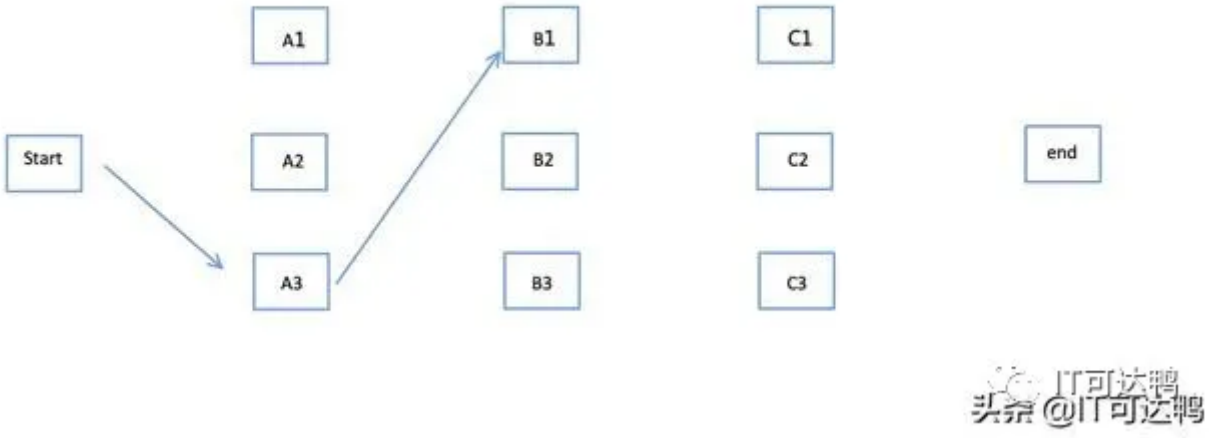
继续往右看，看到了B列。按B列的B1、B2、B3逐个分析。

先看B1:

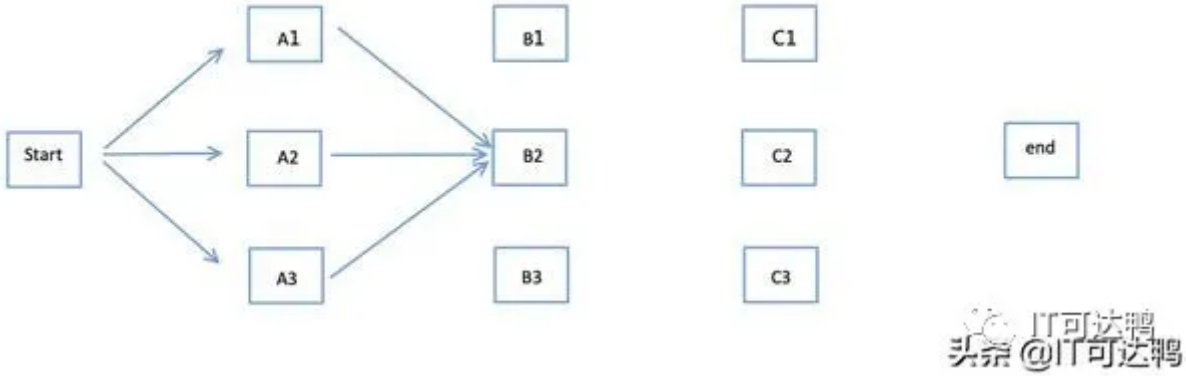


如上图，经过B1的所有路径只有三条：Start-A1-B1、Start-A2-B1、Start-A3-B1。

这三条路径，各个节点的距离加起来对比，就知道其中哪一条最短。假设Start-A3-B1最短，那么其他两条路径就可以大胆的删除。现在所有经过B1的路径只剩下一条路径，如下图:

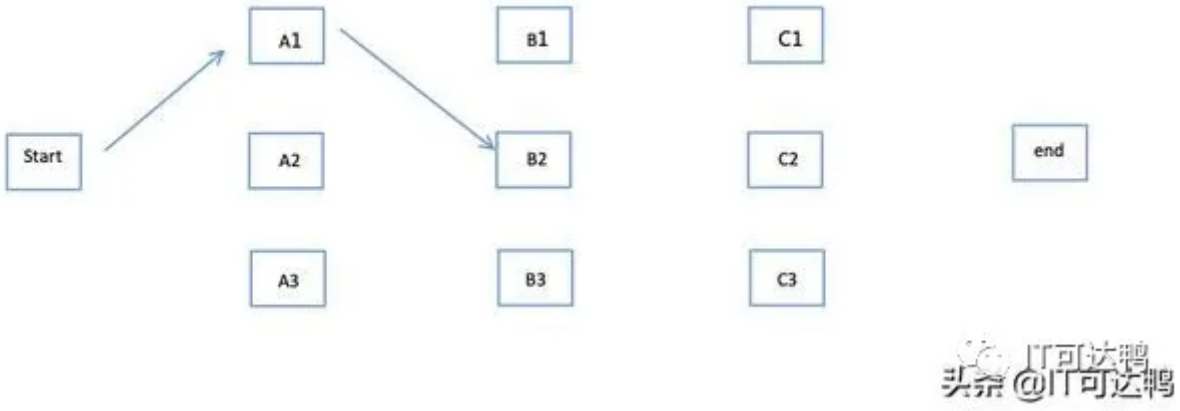


接下来，继续看B2:

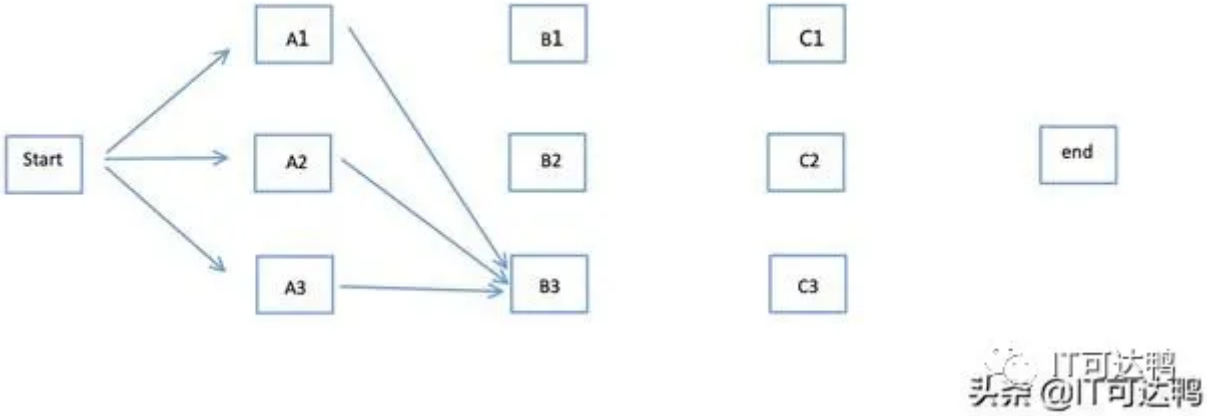


同理，如上图，经过B2的所有路径只有三条：Start-A1-B2、Start-A2-B2、Start-A3-B2。

这三条路径，各个节点的距离加起来对比，就知道其中哪一条最短。假设Start-A1-B2最短，那么其他两条路径就可以大胆的删除。现在所有经过B2的路径只剩下一条路径，如下图：

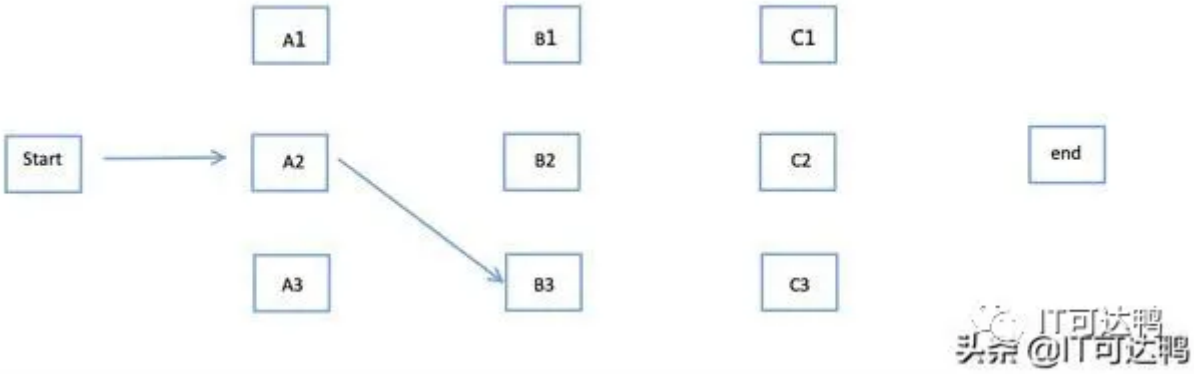


接下来，继续看B3：

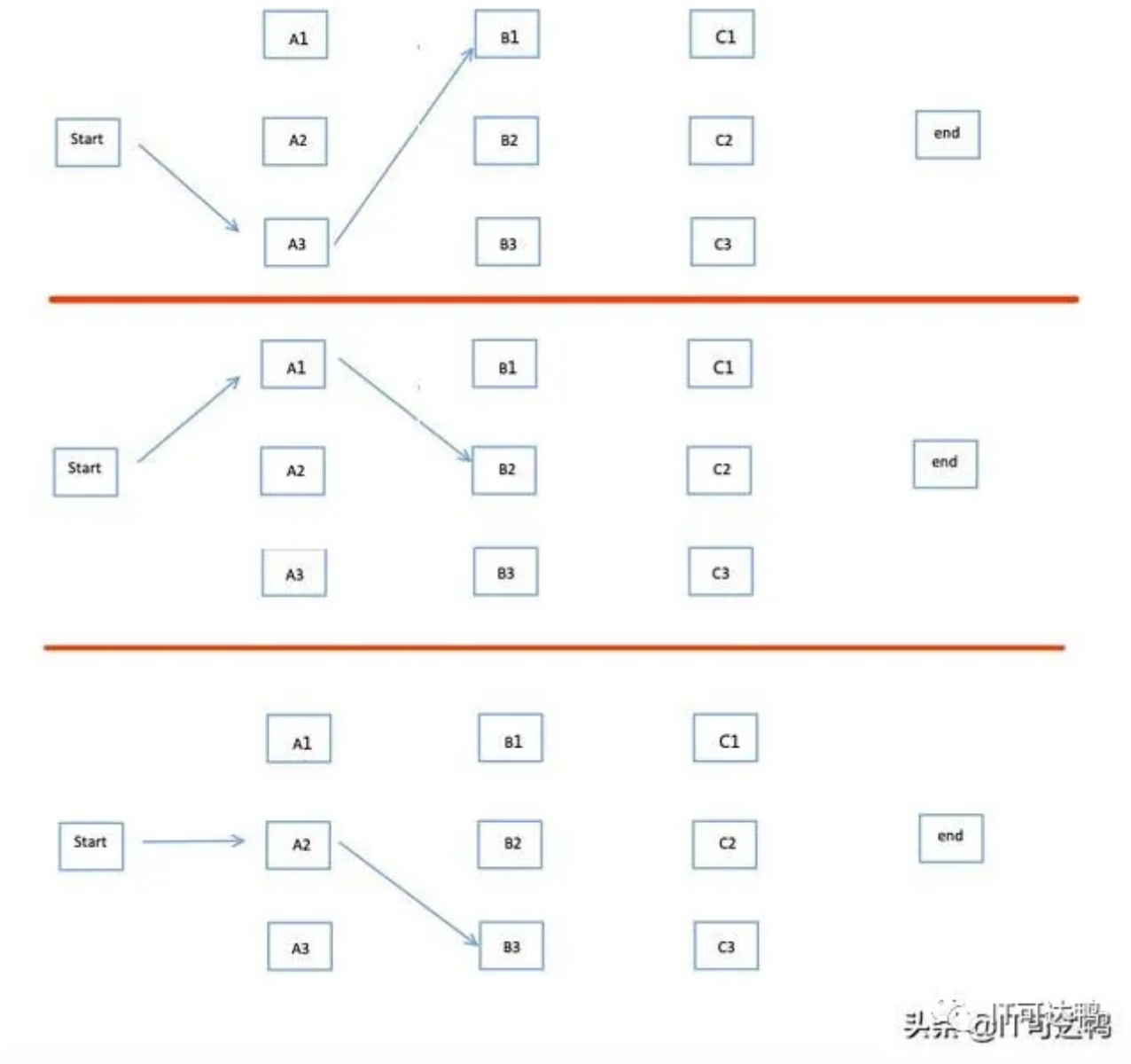


同理，如上图，经过B3的所有路径只有三条：Start-A1-B3、Start-A2-B3、Start-A3-B3。

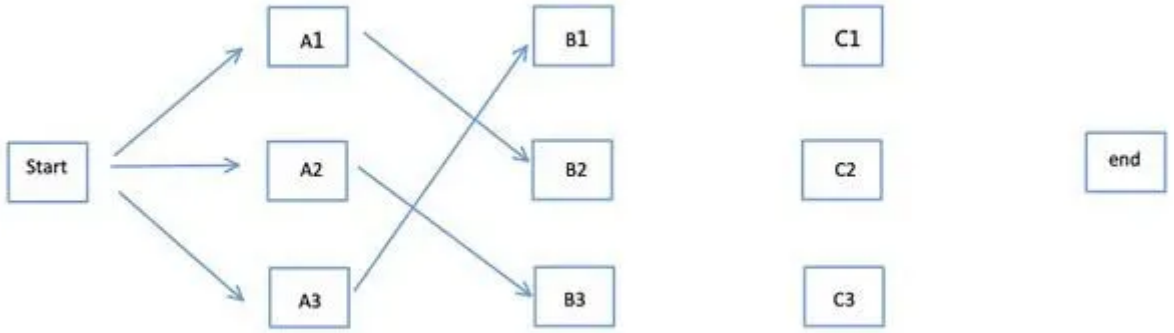
这三条路径，各个节点的距离加起来对比，就知道其中哪一条最短。假设Start-A2-B3最短，那么其他两条路径就可以大胆的删除。现在所有经过B2的路径只剩下一条路径，如下图：



现在对于B列的所有节点我们都过了一遍，B列的每个节点我们都删除了一些不可能是答案的路径，看看我们剩下哪些备选的最短路径，如下图：

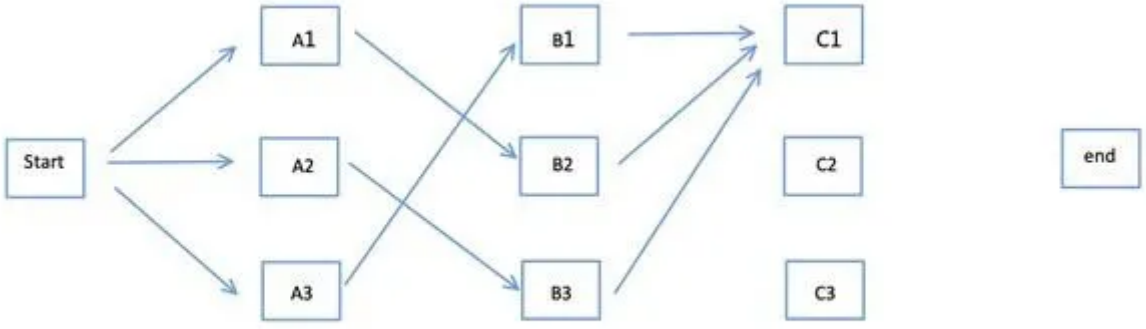


上图是我们删掉了其它不可能是最短路径的情况，留下了三个有可能是最短的路径：Start-A3-B1、Start-A1-B2、Start-A2-B3。现在我们将这三条备选的路径放在一起汇总到下图：

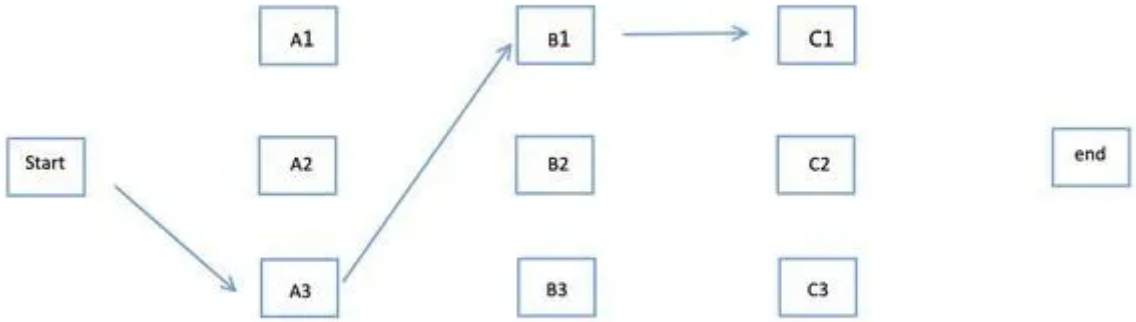


至此，Start-A-B的最优备选路径有3条已经确定，继续往下看C列，这个时候，如果不明白，可以回头再看一遍，前面的步骤决定你是否能看懂viterbi算法。

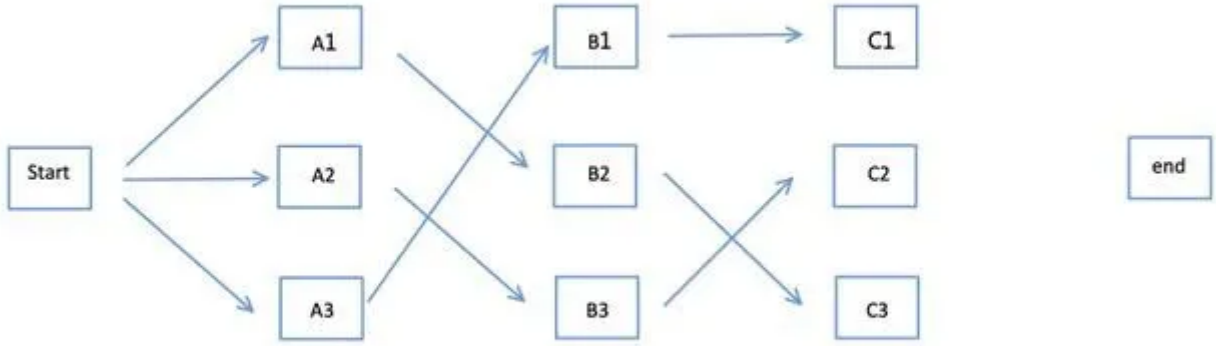
接下来讲C列，从C1、C2、C3一个个节点进行分析，经过C1节点路径有：Start-A3-B1-C1、Start-A1-B2-C1、Start-A2-B3-C1。



和B列做法一样，从这三条路径中找到最短的那条（假定是Start-A3-B1-C1）

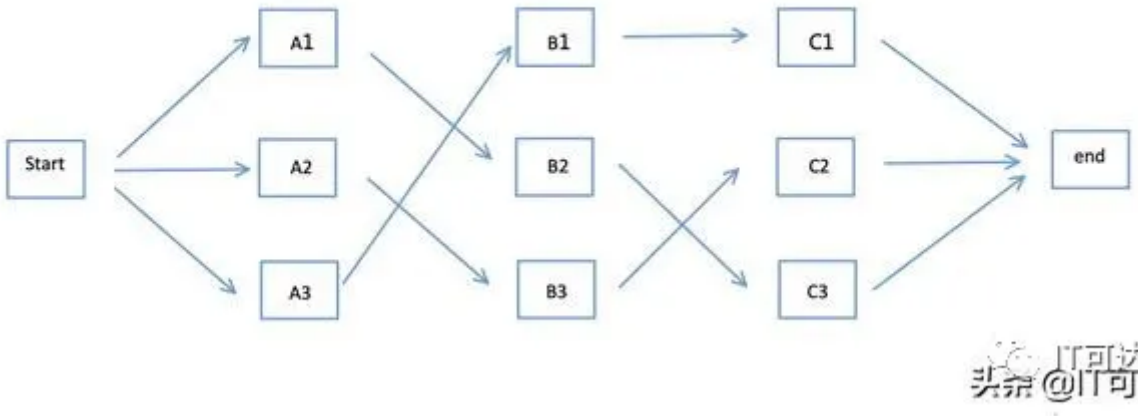


同理，我们可以找到经过C2和C3节点的最短路径，汇总一下：

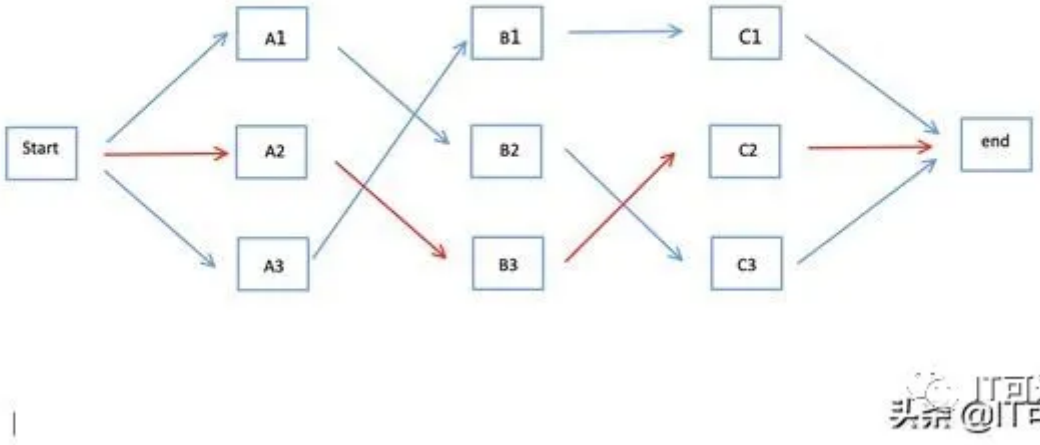


到达C列时最终也只剩3条备选的最短路径，我们仍然没有足够信息断定哪条才是全局最短。最后，我们继续看End节点，才能得出最后的结论。

到End 的路径也只有3种可能性：



E点已经是终点了，我们稍微对比一下这三条路径的总长度就能知道哪条是最短路径了。



所以，对于ABC可观察序列，其标注是2、3、2。

- **结语**

HMM一开始比较难以理解，通过一些简单的例子，可以很好的对HMM的过程进行推理。小编只是通过自己的看法和参考网上的一些例子，对HMM进行整合，难免有不对的地方，欢迎指出。