形式语言与自动机　大作业 1

关于正则语言泵引理的探讨

1951112　林日中

正则语言的泵引理指出，对每一个正则语言 L 都有一个泵长度 p，使得对于该语言中每一个字符串，如果它的长度大于或等于 p 就能够被抽取.

请就下列问题及其它问题进行探讨：

1. 什么是正则语言的泵？为什么需要泵，或者说，泵的作用是什么？

2. 什么是正则语言的泵长度？是否存在最小泵长度？试问：$A = 01^*$，$B = 0^*1^*$，语言 A、B 的泵长度是多少？

3. 同一个 DFA 可能有多个泵吗？如果有多个泵，它的泵长度怎么确定？

4. 如何理解“泵长度 p 是不大于接受正则语言 L 的最小 DFA M 的状态数”？

注意：

1. 本题目是开放性问题，可以对上述问题**及其它问题**进行举例说明、理论证明、探索等.

2. 大作业的成绩计入平时或期末考核范围，具体根据学校期末考核要求决定.

解答.

在形式语言的理论中，泵引理（Pumping Lemma）对于反驳特定语言的正则性很有用. 它在 1959 年由 Michael Rabin 和 Dana Scott 首次证明，不久之后又被 Yehoshua Bar-Hillel、Micha A. Perles 和 Eli Shamir 在 1961 年重新发现，作为他们对上下文无关语言的泵引理的简化.

正则语言的泵引理是描述所有正则语言的基本属性的引理. 非正式地说，它表示可以重复【也即“抽取”（pumping）】常规语言中所有足够长的字符串——也就是说，让字符串的中间部分重复任意次数——以产生一个新的字符串，该字符串也是该语言的一部分.

1. 什么是正则语言的泵？为什么需要泵，或者说，泵的作用是什么？

对于任何正则语言 L，存在一个整数 n，使得 $\forall w \in L, |w| \geq n, \exists x, y, z \in \Sigma^*, w = xyz$，并且

(1) $|xy| \leq n$

(2) $|y| \geq 1$

(3) $\forall i \geq 0, xy^i z \in L$

简单地说，这意味着如果一个字符串 y 是在一个正则语言 L 被“抽取”（“泵”），也就是说，如果 y 被插入一个属于 L 的字符串任意次数，那么产生的字符串仍然保留在 L 中. **我们可以将某个正则语言中共有的重复连续出现的子串 y 称为该正则语言的泵**. 例如，对于正则语言 $L = 0(11)^*101$，泵为子串 11.

泵引理被用来证明语言的非正则性. 因此，如果一种语言是正则的，它总是满足泵引理——如果存在至少一个不在 L 中的由泵产生的字符串，那么 L 肯定是非正则的. 但反之，不一定成立. 也就是说，如果一个语言满足泵引理，并不意味着该语言是正则的.

泵的作用是体现正则语言的构成特征，由某个子串的不同循环重复次数构建正则语言，使得用有限的资源（FA 的状态数、正则表达式的书写长度、正则文法的产生式个数）能够表达无穷集合.

从 FA 的角度，“泵”体现为状态转移路径重叠打圈；从文法的角度，“泵”体现为存在多步推导式 $A \overset{*}{\Rightarrow} yA$，其中，$A$ 为某一变元，y 为正则语言的泵；从正则表达式的角度，“泵”体现为正则表达式中的闭包运算 y^*.

2. 什么是正则语言的泵长度？是否存在最小泵长度？试问：$A = 01^*$，$B = 0^*1^*$，语言 A、B 的泵长度是多少？

由于泵引理是一个蕴含式，正则语言的泵长度是使得泵引理前件成立的串长约束. 注意到，由正则语言的泵引理，一个正则语言 L 的泵为 y，泵长度为 p，$\forall w \in L, |w| \geq p, \exists x, y, z \in \Sigma^*, w = xyz$，其中 $|xy| \leq p$.

如果 p 使得泵引理前件成立，则 $p+1$ 仍能使得泵引理前件成立，因此泵长度没有上界. 记数集 $P = \{p | p \text{ 为泵长度的合法取值}\}$. 由泵引理的证明过程，正则语言 L 对应的 DFA 的状态数可作为 L 的一个泵长度，故有 $P \neq \emptyset$. 由于非空数集 P 有下界（如 $p = 0$ 时，对应的空串无法写为符合泵引理规则的 xyz 的形式），必有下确界，又由于 $P \subseteq \mathbb{N}$，下确界必是最小值，故最小泵长度存在.

对于语言 A，泵长度 $p_A = \{x | x \geq 2, x \in \mathbb{N}\}$，可颐做划分 $x = 0, y = 1^k, z = \varepsilon$，$k \in \mathbb{Z}^+$.

对于语言 B，泵长度 $p_B = \{x | x \geq 1, x \in \mathbb{N}\}$. 对于 B 中的任意串，总可以做划分 $x = \varepsilon, y = 0^k, z = 1^r, k \in \mathbb{Z}^+, r \in \mathbb{N}$ 或 $x = 0^r, y = 1^k, z = \varepsilon, k \in \mathbb{Z}^+, r \in \mathbb{N}$.

为了找到求得正则语言泵长度的通解方法，考虑 DFA M 与正则语言 $L(M)$ 的对应关系. 从初始状态开始，找到一条不形成环的最长路径，其长度加 1 即为最小泵长度 q. 为了找到这样一条路径，我们必须比较从初始状态开始的不同的可能路径，并选择最长的一条. 由于在最长的无环路的长度上加 1 会导致路径中出现环路，对于 $L(M)$ 中达到长度要求的串 w，记其从 M 的起始状态转移至接受状态的过程中，经过的状态序列为 $q_0, q_1, \cdots, q_i, \cdots, q_j, \cdots, q_p$，则存在 $q_i = q_j, 0 \leq i < j \leq q$，故可将 w 拆分为 $q_0 \cdots q_i$、$q_i \cdots q_j$ 和 $q_j \cdots q_p$ 三段，且这三段符合泵引理的要求.

对于有穷语言的泵和泵长度，有穷语言都是正则语言，但不必通过重复子串来表示无限集的结构. 因此，从概念精致性的角度看，有穷语言不应该被假定有泵；但为了与正则语言的性质保持一致，我们规定有穷语言 L 的泵长度是 L 中最长的字符串的长度加 1. 这就定义了有穷语言的泵长度，即有穷语言中没有任何串满足泵长度的前提（$|w| \geq p, w \in L$），从而不会引起矛盾.

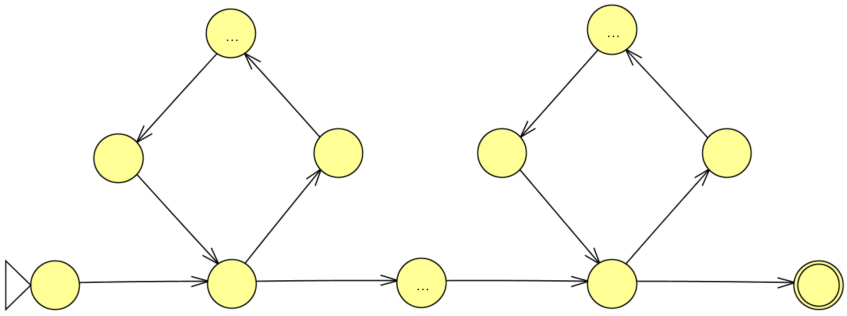
3. 同一个 DFA 可能有多个泵吗？如果有多个泵，它的泵长度怎么确定？

当然有。从泵表示重复多次的子串的定义出发，同一个 DFA 可能有多个泵。为确定其合法的泵长度，只需关注最小泵长度。如果正则语言 L 具有多个泵，最小泵长度仍为**对应**

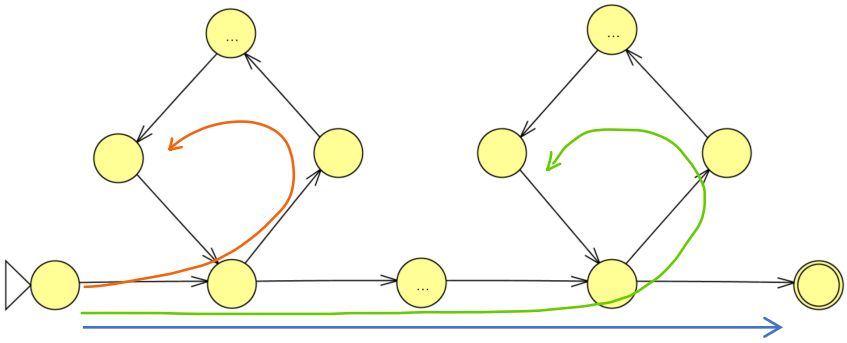
的 DFA 从起始状态开始不会形成环的最长路径的长度加 1。

在搜寻最长路径的过程中，我们比较从起始状态出发的各种可能路径，并从中选出最长的一条。

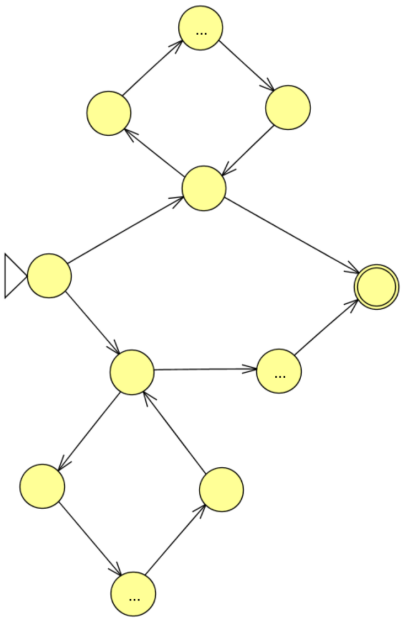
对于连接型的多个泵，被 DFA 接受的串的转移路径如下图所示。



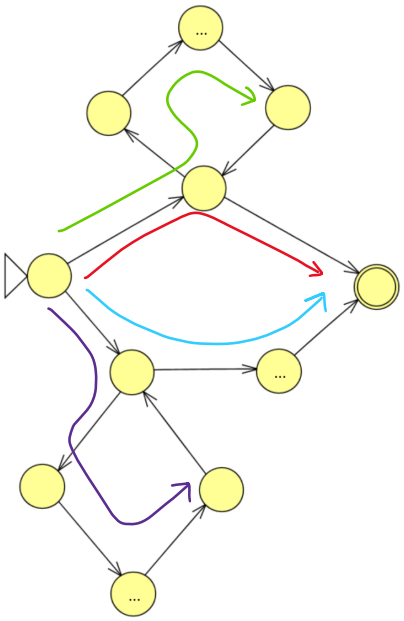
需要比较如下 3 条路径（橙、绿、蓝），选出最长的一条。



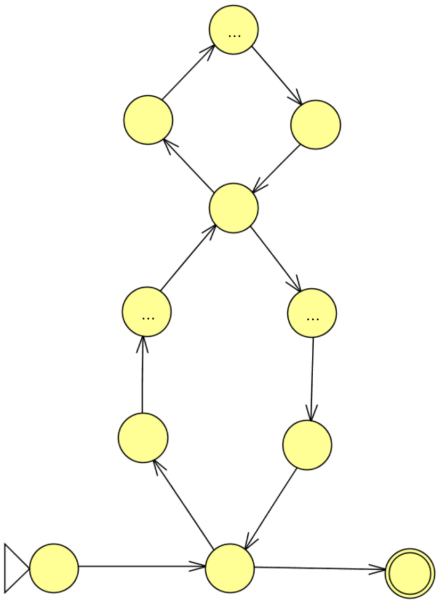
对于并型的多个泵，被 DFA 接受的串的转移路径如下图所示。



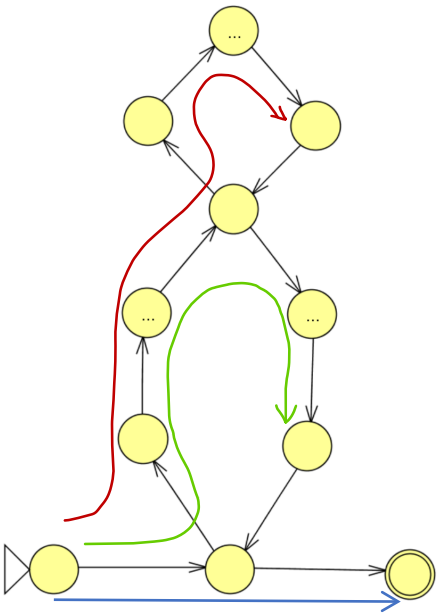
需要比较如下 4 条路径（绿、红、蓝、紫），选出最长的一条。



对于嵌套型的多个泵，被 DFA 接受的串的转移路径如下图所示。



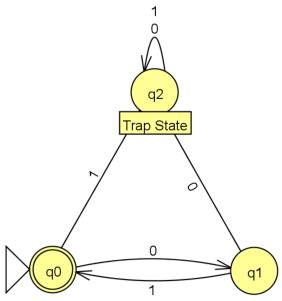
需要比较如下 3 条路径（红、绿、蓝），选出最长的一条。



4. 如何理解“泵长度 p 是不大于接受正则语言 L 的最小 DFA M 的状态数”？

上述陈述中的“泵长度”指的是最小泵长度。根据我们先前对最小泵长度的定义，最小泵长度是对应最短路径的长度加 1，也即最短路径经过的状态数（含首尾）。由于最短路径经过的状态集合是 DFA 状态集合的子集，故最小泵长度 p **不大于**接受正则语言 L 的最小 DFA M 的状态数。

最短路径经过的状态集合是 DFA 状态集合的真子集时，最小泵长度 p **小于**接受正则语言 L 的最小 DFA M 的状态数。例如，正则语言 $L = (01)^*$ 对应的最小 DFA 如下图所示，它包含 3 个状态。然而，L 中的串并不经过陷阱状态 q_2，L 的最小泵长度为 2。



最短路径经过的状态集合与 DFA 状态集合相等时，最小泵长度 p **等于**接受正则语言 L 的最小 DFA M 的状态数。例如，正则语言 $L = 0^*$ 对应的 DFA 如下图所示，它仅包含 1 个状态，而 L 最小泵长度也是 1。