# 正则表达式与自动机

郑奘巍 姜勇刚 11.3

正则表达式出现在20世纪40年代，用来描述正则语言。而到20世纪70年代，它出现在了程序设计领域中。

Michael Fitzgerald

而到20世纪70年代，它出现在了程序设计领域中。

问题1：怎样去判断输入的是否是邮箱地址、用

问题2：怎样在文本中快速定位 计科x班、电计

# 比较

- 字面比较
- 正则表达式是一种字符串模式，用于匹配一组字符串
- 语法（运用层面上的）
- 应用

# 基本语法

- 字符串字面值
- 字符组与选择（元字符）
- 通配符
- 量词
- 位置符
- http://tool.oschina.net/regex#

# 未提及的内容

- 其他语法
  - 字符组缩写、捕获分组、反向引用、环视、量词的贪心与懒惰…
- 内部实现
  - 自动机实现...

# 不同语言的库

- C/C++　　PCRE库/RE库/REEC库
- Python

```
def parsePage(ilt,html):
    try:
        plt = re.findall(r'\"view_price\"\:\"[\d|\.]*\"',html)
        tlt = re.findall(r'\"raw_title\"\:\".*?\"',html)
        for i in range(len(plt)):
            price = eval(plt[i].split(':')[1])
            title = eval(tlt[i].split(':')[1])
            ilt.append([price,title])
    except:
        print("F!")
```

```
序号      价格              商品名称
   1     61.20              灵多种出血牙膏组合牙周牙龈嘴溃疡上火起泡

   2     39.10              人工智能：复杂问题求解的结构和策略（英文版.第6版）
   3     39.10              人工智能：复杂问题求解的结构和策略（英文版.第6版）
   4     75.00              Java程序设计与问题求解(第7版)/萨维切(Walter Savitch)，金名
   5     48.30              正版全新 高等应用数学问题的MATLAB求解(第3版)/薛定宇
   6     68.00              点源函数和边界元法-求解油藏渗流问题
   7     78.80              *正版部分包邮 C++数据抽象和问题求解(第6版)（国外计算机科学经典教材） 编程技术
ADT的主要应用 标准模板库(STL) 软件工程
   8     25.65              创造性问题求解的策略/（美）H.斯科特&middot;福格勒，（美）勒
   9     39.30              工程问题C语言求解 畅销书籍 计算机 正版工程问题C语言求解(原书第4版)
  10     30.50              算法设计与问题求解--编程实践(高等学校规划教材) 书 李清勇 ***工业 正版
```

正则表达式出现在20世纪40年代，用来描述正则语言。而到20世纪70年代，它出现在了程序设计领域中。

Michael Fitzgerald

**正则表达式出现在20世纪40年代，用来描述正则语言。**

- 什么是正则语言
- 正则表达式与状态自动机的关系

# 形式语言

- 专门研究语言的语法的数学和计算机科学分支叫做形式语言理论，它只研究语言的语法而不致力于它的语义。

# 文法

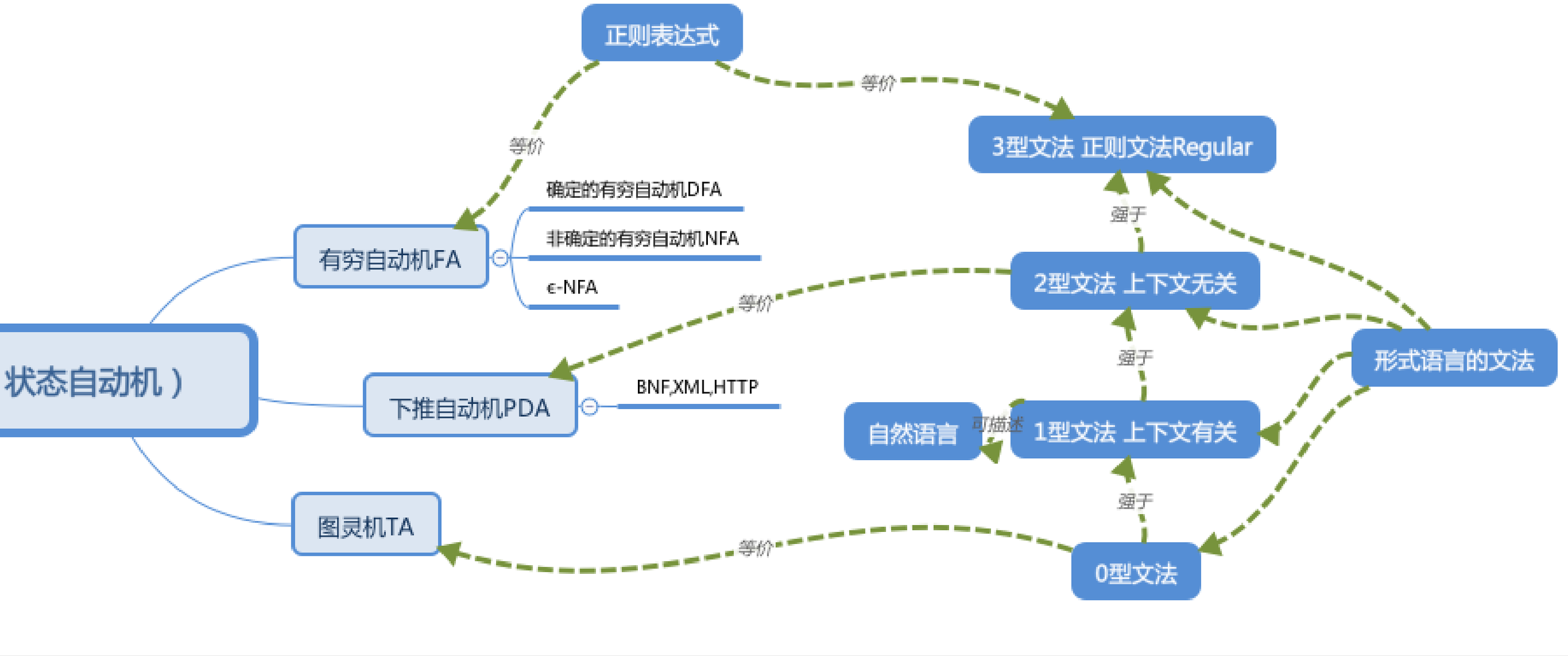
The man eats the apple

A man sings

The cat runs a man

- V = {Sentence, Noun phrase, Article, Noun, Verb phrase, Verb}
- T = {the, a, apple, cat, man, eats, sings, runs}
- S = Sentence
- P = <Sentence> -> <Noun phrase> <Verb phrase>
  - <Noun phrase> -> <Article><Noun>
  - <Article> -> the | a  <Noun> -> apple | cat | man
  - <Verb phrase> -> <Verb><Noun phrase> | <Verb>
  - <Verb> -> ears | sings | runs

# 关系


正则表达式
等价
等价
3型文法 正则文法Regular
强于
确定的有穷自动机DFA
非确定的有穷自动机NFA
ε-NFA
有穷自动机FA
等价
2型文法 上下文无关
强于
形式语言的文法
状态自动机）
下推自动机PDA
BNF,XML,HTTP
自然语言
可描述
1型文法 上下文有关
强于
图灵机TA
等价
0型文法

# 正则文法和正则表达式

- 正则文法
- 正则表达式
- 符号的扩展
- 等价性（3型文法，有穷状态自动机）
- 局限性：只能处理3型文法，最弱

# 参考资料与网站

- 陈有祺 《形式语言与自动机》 P21~P80,P124,P148~P149
- Michael Fitzgerald *Introducing Regular Expressions* *中译本*
- *太极儒的博客* http://blog.sina.com.cn/s/blog_64ac3ab10100ges2.html
- http://www.cnblogs.com/nullzx/p/7092157.html
- http://tool.oschina.net/regex#