Atitit 发帖机系列（7） 词法分析的方法attilax大总结）

[1.1. 词法分析貌似俩大方法，一个直接根据状态图转换，一个根据dfa 1](#_Toc16680)

[1.2. switchcase或者ifelse 最原始方法 1](#_Toc25901)

[1.3. .  状态表 比较实用 2](#_Toc25398)

[1.4.  使用NFA、DFA构建FSM（ 专业方法，难度大） DFA实际上就是高级版的状态表 2](#_Toc24192)

[1.5. 构建词法分析器一般需要几个步骤: 2](#_Toc12644)

[1.5.1. 为正规式设计NFA 由正规式构造FA——Thompson法 2](#_Toc15265)

[1.6.](#_Toc24735) [优先递归 替换循环 3](#_Toc24735)

[1.7. 状态转移表 3](#_Toc305)

界面与后端通讯需要传递dsl，需要做词法分析。。

## 词法分析貌似俩大方法，一个直接根据状态图转换，一个根据dfa

简单点儿说，词法分析就是进行正则表达式匹配。词法分析程序就是根据要匹配的正则表达式生成它的NFA或者DFA，再将待匹配的字符串放到这些NFA或者DFA中进行处理，从而分析出输入字符串是否匹配给定的正则表达式

词法分析器的任务是按照一定模式从源程序中识别出记号(token).

我们使用正规式描述这一模式,并通过有限自动机进行识别.

因为NFA对状态转移不加限制在实际应用中带来很多问题, 通常我们将NFA转换为等价的DFA. 这里所谓的自动机等价是指它们识别同样的正规集.

## [switchcase或者ifelse](http://blog.csdn.net/attilax/article/details/41653319" \l "t11" \t "http://blog.csdn.net/attilax/article/details/_blank) 最原始方法

这无意是最直观的方式，使用一堆条件判断，会编程的人都可以做到，对简单小巧的状态机来说最合适，但是毫无疑问，这样的方式比较原始，对庞大的状态机难以维护。

但checkStateChange()和performStateChange()这两个函数本身依然会在面对很复杂的状态机时，内部逻辑变得异常臃肿，甚至可能是难以实现。

在很长一段时期内，使用switch语 句一直是实现有限状态机的唯一方法，甚至像编译器这样复杂的软件系统，大部分也都直接采用这种实现方式。但 之后随着状态机应用的逐渐深入，构造出来的状态 机越来越复杂，这种方法也开始面临各种严峻的考验，其中最令人头痛的是如果状态机中的状态非常多，或者状 态之间的转换关系异常复杂，那么简单地使用switch语句构造出来的状态机将是不可维护的。

## .  状态表 比较实用

## 使用NFA、DFA构建FSM（ 专业方法，难度大） DFA实际上就是高级版的状态表

使用DFA的方法完成的可配置词法分析器的性能是相当好

一般来说，比较高性能的DFA的实现是一张二维的表。行代表字符，列代表DFA 的状态，单元格代表该状态经输入某个字符之后进行转移的目标状态。此外还有一张表用来记录哪些状态对应哪些规则的结束状态

## 构建词法分析器一般需要几个步骤:

用正规式描述记号的模式

### 为正规式设计NFA [由正规式构造FA——Thompson法](http://www.baidu.com/link?url=xOg_KsebY9ASi3JnrQhpV6TZHolzUoCP5fAFI_ZCcsJGb_z4E5467ZkHNV5-CP7x2ZoW-NI8OdutWCzhhQAV8HW5JQck6lXLkR1bHG270Kq" \t "https://www.baidu.com/_blank)

将NFA转换为等价的DFA, 这一步称为确定化

优化DFA使其状态数最少, 这一步称为最小化

## 优先递归 替换循环

递归可读性更好。。

## 状态转移表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| cur\_dbquo\_stat | 当前状态 | 当前字符 | 要即将转换到的下一状态 |  |
|  |  | \ |  |  |
|  | “ dbQuo\_start | ‘ | <none> |  |
|  | Not sQuo start  Not dbQuo\_start | ‘ | sQuo start |  |
|  | sQuo start | ‘ | sQuo end |  |
| Dbquo end or <non>  Not Dbquo start |  | “ | Dbquo start |  |
| Dbquo start |  | “ | Dbquo end |  |
|  | Non sQuo dbquo start | , |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

字符串优先使用单引号，方便输入。。

meth(\"select from tab where a='abc'\",'str2',\'s3\')

引号需要单独的状态表示

参考资料

atitit.自己动手开发编译器and解释器(1) ------词法分析--attilax总结 - attilax的专栏 - 博客频道 - CSDN.NET.html

词法分析实战 - booirror的博客 - 博客频道 - CSDN.NET.html

现代编译原理－－第一章（词法分析） - BlackWalnut - 博客园.html

作者:: 绰号:老哇的爪子 （ 全名：：Attilax Akbar Al Rapanui 阿提拉克斯 阿克巴 阿尔 拉帕努伊 ）

汉字名：艾提拉（艾龙），   EMAIL:1466519819@qq.com

转载请注明来源： http://blog.csdn.net/attilax

--Atiend