基于网页特征的信息抽取流程图

特征说明>>

* 自身特征：

1. Len(node): node中中文字符的长度。本例中为4. lenThre=30

2. Den(node):= Len(node)/H(node)。 node中中文的密度。H(node)为含标签的整个节点的长度，本文中为13. denThre=0.5

3. Num(node):node中句号和逗号的个数。 numThre=1

* 上下文信息：

4.數據區域適當放鬆條件：如段落title,可能沒有標點，字數也少。但是若他與其他數據區域的父節點是同一個節點，則該節點也很有可能屬於數據區域。具體做法是，當每個節點被判定為數據區域時，則其父節點的權重加1。weiThre=2 （通常同一篇文章有相同的上下文）

