什么是Ansj？

Ansj中文分词是一款纯Java的、主要应用于自然语言处理的、高精度的中文分词工具，目标是“准确、高效、自由地进行中文分词”，可用于人名识别、地名识别、组织机构名识别、多级词性标注、关键词提取、指纹提取等领域，支持行业词典、用户自定义词典。

1. 中文中的歧义

比较出名的一句话“结婚的和尚未结婚的”，如果使用正向最大匹配，容易分成“结婚/的/和尚/未/结婚的”，于是有的学者试图倒过来识别，逆向匹配会大于正向。但是碰到这句“结合成分子时”，采用逆向最大匹配，则会分为“结合/成分/子时”，更有甚者像“咬了猎人的狗”这种语意不明的词语，就更不容易正确分词了。这是中文分词的软肋。下面是些典型的歧义句：

※交叉歧义（多种切分交织在一起）：内塔内亚胡说的/确实/在理

※组合歧义（不同情况下切分不同）：这个人/手上有痣、我们公司人手

※真歧义（几种切分都可以）：乒乓球拍/卖/完了、乒乓球/拍卖/完了

1. 实体名识别

这个是中文分词遇到的最大的难点，也是最最紧迫的。实体名识别包括人名识别、地名识别、机构名识别，还包括有监督识别和无监督识别。有监督的还好，无监督基本是无解的，比如“王大力发球”是“王大力”还是“大力发球”，一般人都难以识别。

1. 新词热词发现

目前常用的新词发现还是一个比较有研究性的课题，虽然有些论文在准确率很高，但是大多是封闭测试，这意味着结果很难应用到实际工程中。目前Ansj采用的新词发现方式比较简单，采用了高频词的匹配方式，不使用规则，用统计重复串识别新词，根据词性去掉干扰词，虽然有一定的效果，但还是差强人意。

1. 颗粒度问题

比如“北京大学”是“北京”+“大学”还是“北京大学”，人各有志，就连同一个人不同时间的标注也有可能是有区别的，虽然这个问题严格上来说不属于技术问题，但是对分词结果的评测却有着很大的关系，Ansj采用“能识别就识别”的策略方针，所以在真正R值的时候偏低，总之一句话，适合学术的不一定适合工业，反之亦然。

Ansj的算法原理

该算法实现分词有以下几个步骤：

1、 全切分，原子切分；

2、 N最短路径的粗切分，根据隐马尔科夫模型和viterbi算法，达到最优路径的规划；

3、人名识别；

4、 系统词典补充；  
5、 用户自定义词典的补充；  
6、 词性标注（可选）

以上转自博客：开源Java中文分词器Ansj作者孙健专访。

博客地址：<http://blog.csdn.net/blogdevteam/article/details/8148451>

尚未完成的部分：

1. 增加一些规则Recognition，举例身份证号码识别，目前未完成的有 时间识别，ip地址识别，邮箱识别，网址识别，词性识别。