# 华为相似语料抽取方法

刘毅 <liuyi@torangetek.com>

2014 年 2 月 12 日

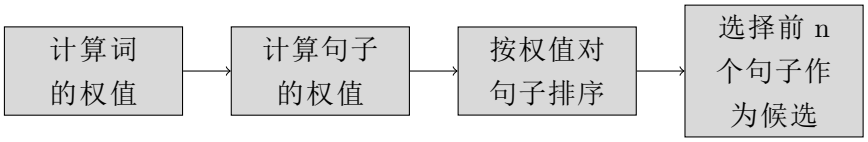
## 1 需求分析

项目需要在新闻语料库和专利语料库中抽取 500 万行左右的句子，要求抽取的句子与华为语料库中的句子相似，如使用了相同的词语，相同的语法，等特征。

## 2 基本方法和原理

我们对新闻和专利语料库中的句子进行评分，然后根据分数进行排序，抽取分数较高者。

### 2.1 整体流程




### 2.2 分数的定义

我们对分数的定义有如下需求：

1. 被抽取的句子中的词在华为语料库中反复出现，频率较高。
2. 若源语料库中（新闻，专利）的某个句子中所有的词汇从未来华为语料库中出现过，那么该句子的得分为 0。若源语料库中的某个句子与华为语料库中的某句子完全相同，那么该句子应该得到相当高的分数。

3. 若源语料库中的某一句话的各个词汇分布在华为语料库中不同的句子中，该句子的分数也应该比较高。