信息检索导论--
中文分词技术

授课人：李永可
邮箱：lyk@xjau.edu.cn

中文分词技术

中文分词

- 中文分词 (Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

中文分词

- 中文分词是文本挖掘的基础，对于输入的一段中文，成功的进行中文分词，可以达到电脑自动识别语句含义的效果。

中文分词

- 中文分词技术属于自然语言处理技术范畴，对于一句话，人可以通过自己的知识来明白哪些是词，哪些不是词，但如何让计算机也能理解？其处理过程就是分词算法。

- 例如：茶和服装
- 分词结果:

（1）茶|和|服装

（2）茶|和服|装

常用中文分词算法

- 现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。

基于字符串匹配的分词方法

- 这种方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。按照扫描方向的不同，字符串匹配分词方法可以分为正向匹配和逆向匹配；按照不同长度优先匹配的情况，可以分为最大（最长）匹配和最小（最短）匹配；常用的几种机械分词方法如下：

基于字符串匹配的分词方法

1）正向最大匹配法（由左到右的方向）；

2）逆向最大匹配法（由右到左的方向）；

3）最少切分（使每一句中切出的词数最小）；

4）双向最大匹配法（进行由左到右、由右到左两次扫描）

分词算法遵循几个原则

- 颗粒度越大越好：

用于进行语义分析的文本分词，要求分词结果的颗粒度越大，即单词的字数越多，所能表示的含义越确切，如：“公安局长”可以分为“公安 局长”、“公安局 长”、“公安局长”都算对，但是要用于语义分析，则“公安局长”的分词结果最好（当然前提是所使用的词典中有这个词）

分词算法遵循几个原则

- 切分结果中非词典词越少越好：

单字字典词数越少越好，这里的“非词典词”就是不包含在词典中的单字，而“单字字典词”指的是可以独立运用的单字，如“的”、“了”、“和”、“你”、“我”、“他”。例如：“技术和服务”，可以分为“技术 和服 务”以及“技术 和 服务”，但“务”字无法独立成词（即词典中没有），但“和”字可以单独成词（词典中要包含），因此“技术 和服 务”有1个非词典词，而“技术 和 服务”有0个非词典词，因此选用后者

分词算法遵循几个原则

- 总体词数越少越好：

在相同字数的情况下，总词数越少，说明语义单元越少，那么相对的单个语义单元的权重会越大，因此准确性会越高。

正向最大匹配算法

最大匹配是指以词典为依据，取词典中最长单词长度Maxlen为第一个次取字数量的扫描串，在词典中进行扫描，如果在词典中发现匹配字符串，则匹配成功，当前字符串即为分词结果，若没有匹配字符串，则将当前处理字符串最右侧字去掉，然后重新开始一轮匹配，直到找到匹配字符串为至，若没有一个字符串能够匹配，则截取待匹配字符串第一个字符为匹配结果。

正向最大匹配算法

- 以“我们在野生动物园玩”详细说明一下正向最大匹配方法：

正向即从前往后取词，从7->1，每次减一个字，直到词典命中或剩下1个单字。

第1次：“我们在野生动物”，扫描词典，无

第2次：“我们在野生动”，扫描词典，无

。。。。

第6次：“我们”，扫描词典，有

扫描中止，输出第1个词为“我们”，去除第1个词后开始第2轮扫描

正向最大匹配算法

- 第2轮扫描：
- 第1次：“在野生动物园玩”，扫描词典，无
- 第2次：“在野生动物园”，扫描词典，无
- 。 。 。 。
- 第6次：“在野”，扫描词典，有
- 扫描中止，输出第2个词为“在野”，去除第2个词后开始第3轮扫描

正向最大匹配算法

- 第3轮扫描：
- 第1次：“生动物园玩”，扫描词典，无
- 第2次：“生动物园”，扫描词典，无
- 第3次：“生动物”，扫描词典，无
- 第4次：“生动”，扫描词典，有
- 扫描中止，输出第3个词为“生动”，第4轮扫描

正向最大匹配算法

- 第4轮扫描：
- 第1次：“物园玩”，扫描词典，无
- 第2次：“物园”，扫描词典，无
- 第3次：“物”，扫描词典，无
- 扫描中止，输出第4个词为“物”，非字典词数加1，开始第5轮扫描

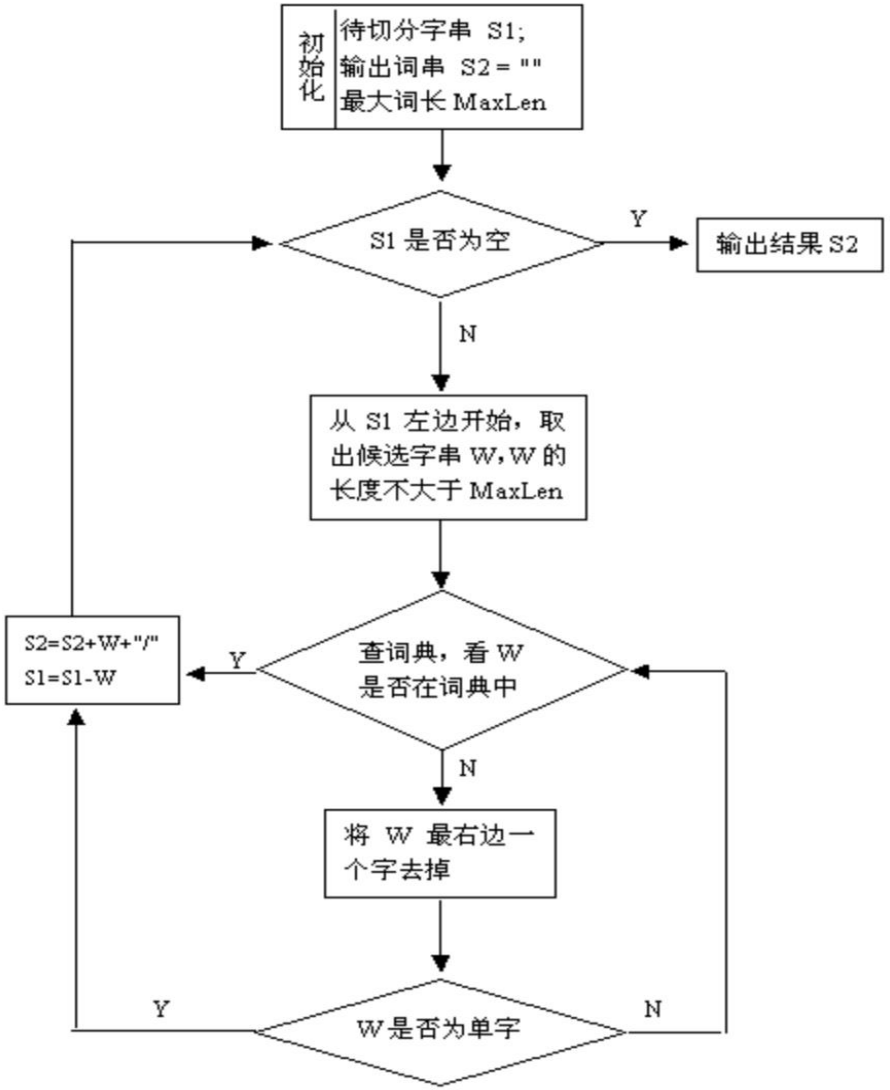
正向最大匹配算法

- 第5轮扫描：
- 第1次：“园玩”，扫描词典，无
- 第2次：“园”，扫描词典，有
- 扫描中止，输出第5个词为“园”，单字字典词数加1，开始第6轮扫描

正向最大匹配算法

- 第6轮扫描：
- 第1次：“玩”，扫描字典，有
- 扫描中止，输出第6个词为“玩”，单字字典词数加1，整体扫描结束。
- 正向最大匹配法，最终切分结果为：“我们/在野/生动/物/园/玩”，其中，单字字典词为2，非词典词为1。

正向最大匹配算法


初始化
待切分字串 S1;
输出词串 S2 = ""
最大词长 MaxLen
S1 是否为空
Y
输出结果 S2
N
从 S1 左边开始，取出候选字串 W，W 的长度不大于 MaxLen
查词典，看 W 是否在词典中
Y
S2=S2+W+"/"
S1=S1-W
N
将 W 最右边一个字去掉
W 是否为单字
Y
N

正向最大匹配算法

- 参考网页：
- http://yangshangchuan.iteye.com/blog/2031813

逆向最大匹配算法

- 逆向即从后往前取词，其他逻辑和正向相同。即：
- 第1轮扫描：“在野生动物园玩”
- 第1次：“在野生动物园玩”，扫描词典，无
- 第2次：“野生动物园玩”，扫描词典，无
- 。 。 。 。
- 第7次：“玩”，扫描1字词典，有
- 扫描中止，输出“玩”，单字字典词加1，开始第2轮扫描

逆向最大匹配算法

- 第2轮扫描：“们在野生动物园”
- 第1次：“们在野生动物园”，扫描词典，无
- 第2次：“在野生动物园”，扫描词典，无
- 第3次：“野生动物园”，扫描词典，有
- 扫描中止，输出“野生动物园”，开始第3轮扫描

逆向最大匹配算法

- 第3轮扫描：“我们在”
- 第1次：“我们在”，扫描3字词典，无
- 第2次：“们在”，扫描2字词典，无
- 第3次：“在”，扫描1字词典，有
- 扫描中止，输出“在”，单字字典词加1，开始第4轮扫描

逆向最大匹配算法

- 第4轮扫描：“我们”
- 第1次：“我们”，扫描2字词典，有
- 扫描中止，输出“我们”，整体扫描结束。
- 逆向最大匹配法，最终切分结果为：“我们/在/野生动物园/玩”，其中，单字字典词为2，非词典词为0。

基于理解的分词方法

- 这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

统计法

- 从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。定义两个字的互现信息，计算两个汉字X、Y的相邻共现概率。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词。

技术难点

- 到底哪种分词算法的准确度更高，目前并无定论。对于任何一个成熟的分词系统来说，不可能单独依靠某一种算法来实现，都需要综合不同的算法。

技术难点

- 有了成熟的分词算法，是否就能容易的解决中文分词的问题呢？事实远非如此。中文是一种十分复杂的语言，让计算机理解中文语言更是困难。在中文分词过程中，有两大难题一直没有完全突破。

技术难点--歧义识别

- 歧义是指同样的一句话，可能有两种或者更多的切分方法。主要的歧义有两种：交集型歧义和组合型歧义，例如：表面的，因为“表面”和“面的”都是词，那么这个短语就可以分成“表面 |的”和“表| 面的”。这种称为交集型歧义（交叉歧义）。

技术难点--歧义识别

- 交集型歧义相对组合型歧义来说是还算比较容易处理，组合型歧义就必须根据整个句子来判断了。例如，在句子“这个门把手坏了”中，“把手”是个词，但在句子“请把手拿开”中，“把手”就不是一个词；在句子“将军任命了一名中将”中，“中将”是个词，但在句子“产量三年中将增长两倍”中，“中将”就不再是词。这些词计算机又如何去识别?

技术难点--歧义识别

- 如果交集型歧义和组合型歧义计算机都能解决的话，在歧义中还有一个难题，是真歧义。真歧义意思是给出一句话，由人去判断也不知道哪个应该是词，哪个应该不是词。例如：“乒乓球拍卖完了”，可以切分成“乒乓 球拍 卖 完 了”、也可切分成“乒乓球 拍卖 完 了”，如果没有上下文其他的句子，恐怕谁也不知道“拍卖”在这里算不算一个词。

技术难点--新词识别

- 命名实体（人名、地名）、新词，专业术语称为未登录词。也就是那些在分词词典中没有收录，但又确实能称为词的那些词。最典型的是人名，人可以很容易理解。句子“王军虎去广州了”中，“王军虎”是个词，因为是一个人的名字，但要是让计算机去识别就困难了。如果把“王军虎”做为一个词收录到字典中去，全世界有那么多名字，而且每时每刻都有新增的人名，收录这些人名本身就是一项既不划算又巨大的工程。即使这项工作可以完成，还是会存在问题，例如：在句子“王军虎头虎脑的”中，“王军虎”还能不能算词？

技术难点--新词识别

- 除了人名以外，还有机构名、地名、产品名、商标名、简称、省略语等都是很难处理的问题，而且这些又正好是人们经常使用的词，因此对于搜索引擎来说，分词系统中的新词识别十分重要。新词识别准确率已经成为评价一个分词系统好坏的重要标志之一。

中文分词应用

- 在自然语言处理技术中，中文处理技术比西文处理技术要落后很大一段距离，许多西文的处理方法中文不能直接采用，就是因为中文必需有分词这道工序。中文分词是其他中文信息处理的基础，搜索引擎只是中文分词的一个应用。其他的比如机器翻译（MT）、语音合成、自动分类、自动摘要、自动校对等等，都需要用到分词。

常用中文分词器

- ICTCLAS
- 这是最早的中文开源分词项目之一，ICTCLAS在国内973专家组组织的评测中活动获得了第一名，在第一届国际中文处理研究机构SigHan组织的评测中都获得了多项第一名。ICTCLAS3.0分词速度单机996KB/s，分词精度98.45%，API不超过200KB，各种词典数据压缩后不到3M.ICTCLAS全部采用C/C++编写，支持Linux、FreeBSD及Windows系列操作系统，支持C/C++、C#、Delphi、Java等主流的开发语言。

常用中文分词器

- IKAnalyzer
- IKAnalyzer是一个开源的，基于java语言开发的轻量级的中文分词工具包。从2006年12月推出1.0版开始，IKAnalyzer已经推出了3个大版本。最初，它是以开源项目Luence为应用主体的，结合词典分词和文法分析算法的中文分词组件。新版本的IKAnalyzer3.0则发展为面向Java的公用分词组件，独立于Lucene项目，同时提供了对Lucene的默认优化实现。

常用中文分词器

- **Paoding**（庖丁解牛分词）基于Java的开源中文分词组件，提供lucene和solr 接口，具有极 高效率和 高扩展性。采用完全的面向对象设计，构思先进。
- 高效率：在PIII 1G内存个人机器上，**1秒**可准确分词 **100万**汉字。
- 采用基于 不限制个数的词典文件对文章进行有效切分，使能够对词汇分类定义。
- 能够对未知的词汇进行合理解析。
- 仅支持Java语言。

其它中文分词

- 盘古分词
- SCWS
- FudanNLP
- HTTPCWS
- CC-CEDICT
- MMSEG4J
- Jcseg

停用词

- 在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。但是，并没有一个明确的停用词表能够适用于所有的工具。甚至有一些工具是明确地避免使用停用词来支持短语搜索的