Ch 6

KNN k-Nearest Neighbor k近邻算法 基于统计

HAC 层次聚类法

K-means k-均值聚类 快速聚类

基本概念

自动标引

1. 用机器抽取或赋予索引词，一旦编制好程序和规则，就不需要人工干预
2. ~~意义~~
    a. 适应信息资源快速增长的需要
    b. 效率高、成本低
    c. 稳定性好、一致性好
3. 自动化程度：全自动标引、半自动标引
4. 标引词的来源：（自动）抽词标引、（自动）赋词标引

（自动）抽词标引

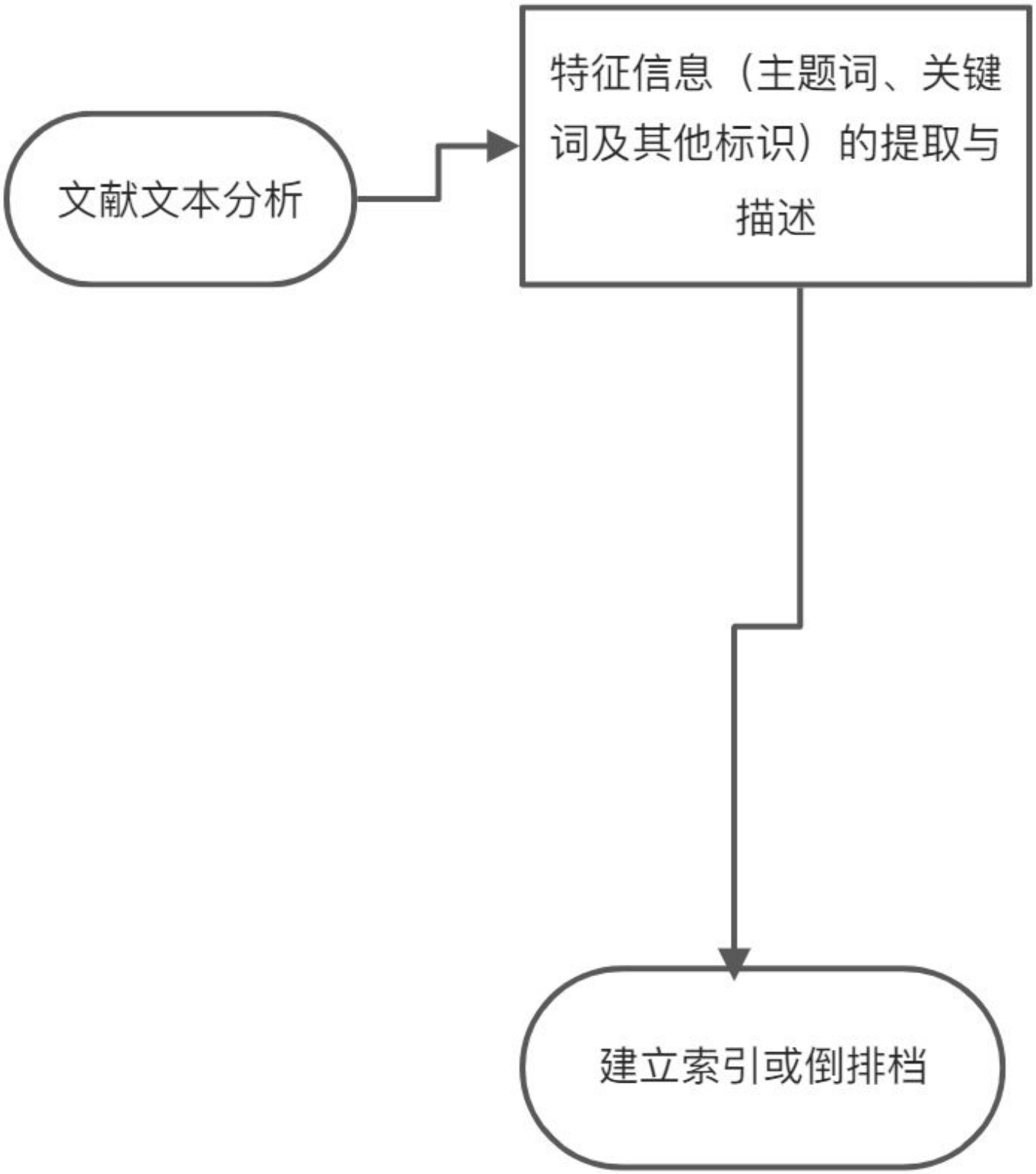
1. 利用计算机直接从文献题名、文摘或正文中自动抽出能表达文献主题的词作为标引词，并自动生成关键词索引或倒排档
2. 抽词标引的标引词只能来源于文献本身的文内关键词，所以也称为自由词标引

（自动）赋词标引

1. 让计算机模仿人的赋词标引方法，分析文献的内容，选取与文献主题相符或密切相关的语词符号作为索引词
2. 其标引词是由描述词组成的，这些词不一定来源于文献本身所用的词，而是选自预先编制的词表，所以叫受控词标引

文献标引

1. 对所收集的文献给出标识导引，这些标识包括文献标题、作者名、分类号和主题词等



~~分词~~

1. 正确的机器自动分词是正确的中文信息处理的基础

切分歧义

1. 歧义类型
    a. 交集型歧义：对于汉字串AJB，AJ、JB同时成词
    b. 组合型歧义：对于汉字串AB，A、B、AB同时成词
    c. 混合型歧义：同时包含交集型歧义和组合型歧义
2. 另一种分类
    a. 真歧义：歧义字段在不同的语境中确实有多种切分形式
    b. 伪歧义：歧义字段单独拿出来看有歧义，但在所有真实语境中，仅有一种切分形式可接受
    c. 对于交集型歧义字段，真实文本中伪歧义现象远多于真歧义现象

未登录词

1. 虽然一般的词典都能覆盖大多数的词语，但有相当一部分的词语不可能穷尽地收入系统词典中，这些词语称为未登录词或新词
2. 专有名词、重叠词、口语、派生词、与领域相关的术语

停用词

1. 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
2. 停用词表是一种特殊的词表，在这个词表中含有冠词、虚词、叹词、连词、介词以及语义泛泛的词等一切在上下文中没有检索意义的词

Zipf定律

1. 如果把一篇较长文章（>5000）中每个词出现的频率统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然语言给这些词编上等级序号，即频次最高的词的等级为1，频次次高的等级为2，……，频次最小的词等级为D（或L），若用f表示等级为r的词在文献中出现的相对频次，则有：$f_r \cdot r = C$ （C是一个常数，大约等于0.1）
2. 齐普夫分布曲线：如果用横坐标表示词的等级序号r，纵坐标表示相应的频次fr，我们就可以得到一条双曲线

自动分类

1. 广义，自动聚类

a. 从待分类对象中提出特征，然后将提出的全部特征进行比较，再根据一定的原则将具有相同或相近特征的对象定义为一类，并设法使各类中包含的对象大致相等
b. 先有文档后有类
c. 无监督学习

2. 狭义，自动归类
   a. 在给定的分类体系下，分析被分类对象的特征，使之与各种类别中对象所具有的共同特征进行比较，然后将对象划归为特征最接近的一类并赋予相应的分类号
   b. 先有类（表）后有文档
   c. 有监督学习

纯度 （恢复人工分类能力的评测及其指标）

1. 聚类效果的评价指标
2. 正确的个数/总文档数

基本流程

特征选择

1. 在文本分类问题中遇到的一个主要困难就是高维的特征空间
   a. 一份普通的文本在经过文本表示后，如果以字/词为特征，它的特征空间维数将达到几千，甚至几万
   b. 大多数学习算法都无法处理如此大的维数
2. 在不牺牲分类质量的前提下尽可能降低特征空间的维数
3. 特征选取的任务将信息量小，不重要的词汇从特征空间中删除，减少特征项的个数
4. 在许多文本分类系统的实现中都引入了特征提取方法
5. 对每一类构造k 个最有区别能力的term

DF Document Frequency

1. DF小于某个阈值的去掉，没有代表性
2. DF大于某个阈值的去掉，没有区分度
3. 优点：降低向量计算的复杂度，去掉部分噪声，提高分类的准确率，且简单易行

4. 缺点：稀少的词具有更多的信息，因此不宜用DF大幅度地删除词

IG Information Gain 信息增益

1. 某term为整个分类所能提供的信息量，即不考虑某特征的熵和考虑该特征的熵的差值
    a. 计算不含任何特征整个文档的熵
    b. 计算包含该特征的文档的熵
    c. 前者-后者，选择Top K作为特征
2. 优点：准确，选择的特征是对分类有用的特征
3. 缺点：有些信息增益较高的特征出现的频率较低

CHI x^2

1. CHI衡量的是特征项t(i)和类C(j)之间的相关联程度
2. 假设t(i)和C(j)之间符合具有一阶自由度的卡方分布，如果特征对于某类的卡方统计值越高，它与该类之间的相关性越大，携带的信息越多，反之则越少
3. 特点：只统计文档是否出现词，而不管出现了几次；夸大了低频词的作用

MI Mutual Information

1. 互信息越大，则特征t(i)和类C(j)之间共同出现的程度越大 ，如果两者无关，那么互信息=0
2. 优点：如果某个特征词的频率很低，那么互信息得分就会很大，因此互信息法倾向"低频"的特征词
3. 缺点：相对的词频很高的词，得分就会变低，如果这词携带了很高的信息量，互信息法就会变得低效

分类评测

	属于此类	不属于此类
判定属于此类	a	b
判定不属于此类	c	d

1. 准确率(precision) = a/(a + b)
2. 召回率(recall) = a/(a + c)

类间相似度

1. 单链 `Single Link` ： 两个聚类间最相似文档的相似度来表示聚类相似度
2. 全链 `Complete Link` ： 两个聚类间最不相似文档的相似度来表示聚类相似度
3. 组平均 `Group Average` ： 两个聚类间文档的平均相似度来表示聚类相似度
4. ~~聚类中心点~~：用中心向量表示聚类，聚类间相似度采用向量夹角余弦

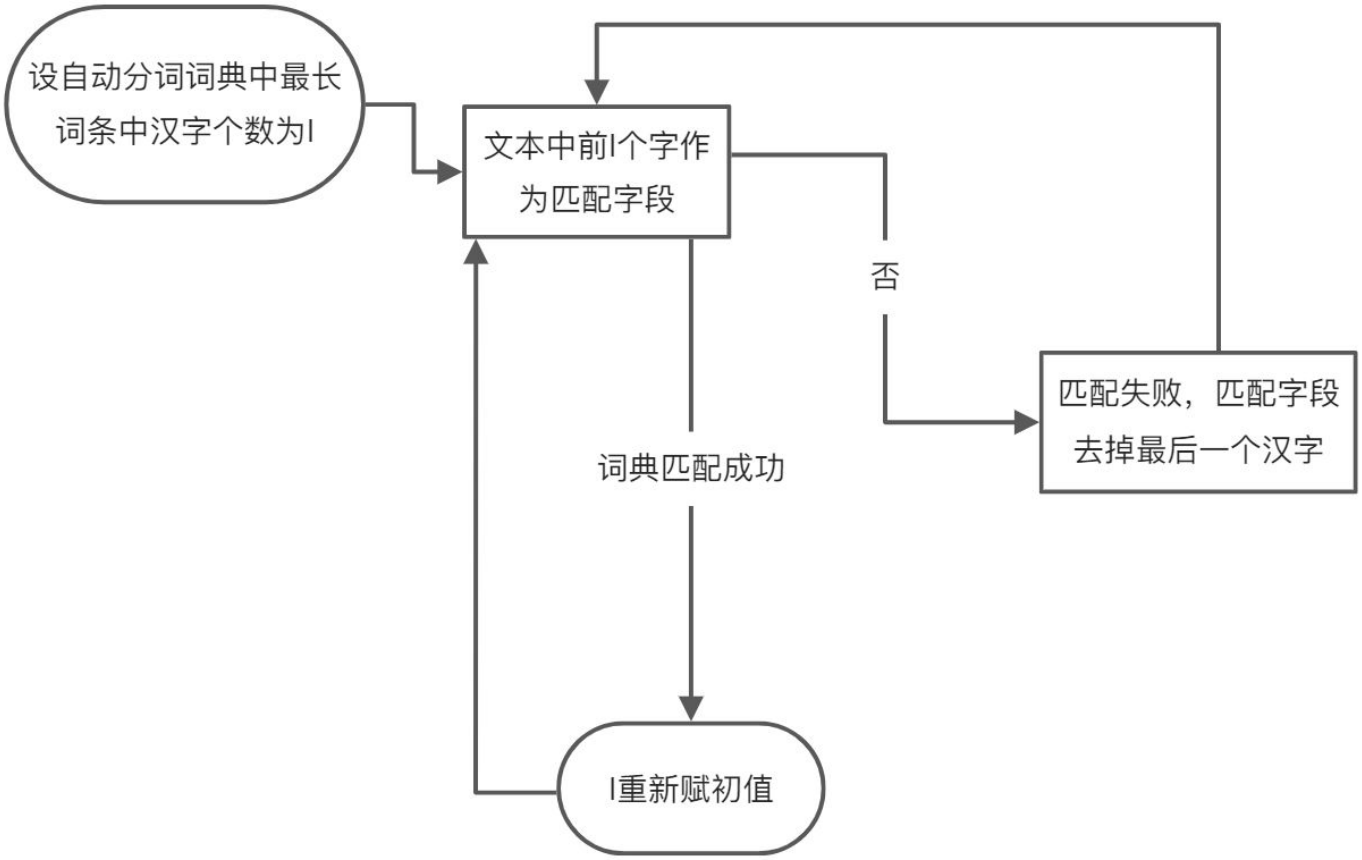
基本流程

FMM 正向最大匹配 `Forward Maximum Matching Method`

特点

1. 对交叉歧义和组合歧义没有什么好的解决办法
2. 往往不单独使用，而是与其它方法配合使用

流程



BMM 逆向最大匹配 Backward Maximum Matching Method

流程

1. 分词过程与FMM方法相同，不过是从句子(或文章)末尾开始处理，每次匹配不成功时去掉的是前面的一个汉字

特点

1. 逆向最大匹配法比正向最大匹配法更有效

分词规则

1. 颗粒度越大越好——词长
2. 非词典词越少越好
3. 总体词数越少越好

KNN k-Nearest Neighbor k近邻算法 基于统计

1. 给定一个经过分类的训练文档集合，在对新文档（即测试文档或待分类文档）进行分类时，首先从训练文档集合中找出与测试文档最相关的k篇文档，然后按照这k篇文档所属的类别信息来对该测试文档进行分类处理

HAC 层次聚类法

1. 一种可以利用谱系结构或树状结构图来描绘聚类过程的方法，也是进行聚类分析时应用最多的方法
2. 分解法：在聚类开始时，将所有的文献都看成是一类，然后再根据距离或相似性，不断进行分解，直到每篇文献都自成一类为止
3. 凝聚法：聚类开始将每篇文献看成一类，然后再根据距离或者相似性，不断进行合并，直到将所有文献都归结为一类为止

K-means k-均值聚类 快速聚类

1. 先对所要分类的事物作一个初始的分类，然后按照某种最优的原则修改不合理的初始分类，直至分类被认为比较合理时为止，形成最终的聚类结果