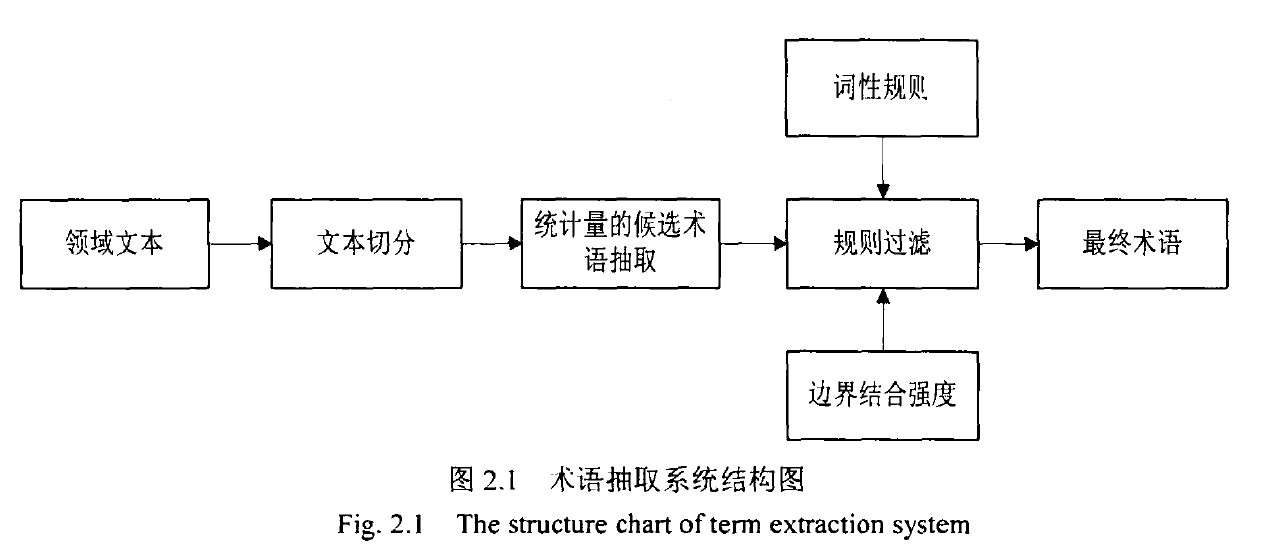
术语抽取系统主要由三部分组成：文本切分，基于词频分布和信息熵的候选术语抽取，规则过滤。



**1、分词**

用<http://ictclas.nlpir.org/>产品，进行分词和词性标注。

**2、建立索引结构。**

应用[patricia tree](http://www.baidu.com/link?url=iZbQ7vNOyuB1Pzw4Rl0FUMzvw5J6NL7pMNH-Ta53NsMs1vOPE8dgZ4qsFR2H4HjKhHFFzORD1Zqf4Vr4kJoTqK" \t "_blank)作为索引结构，对每一个文档建立Pat-tree，用于统计字符串的频率信息。

**3、候选术语抽取**

**①词频分布变化方法**

由于术语具有领域相关性，一个术语在**领域**间分布不均匀，即一个术语在相关领域内出现频率较高，而在不相关邻域内很少出现或几乎不出现。方差的值越小，表明候选术语在文档间的频率变化越小，是术语的可能性也越小。

**词频**分布公式如下所示:





其中，s表示候选术语，表示s在整个语料中的出现频率，表示s出现的文档频率，表示s在第i篇文档中的出现频率，表示在各个文档中出现的平均频率，为总的文档数，表示包含该术语的文档数。

**②信息熵方法**

信息论中的信息熵表示单个随机变量的不确定性。随机变量越不确定，其熵值越大。当信息熵用于术语抽取时，主要用于计算字符串的边界不确定性。字符串的边界越不确定，信息熵越高，则越可能是一个完整的词。

一般通过计算字符串的左信息熵和右信息熵来衡量字符串的左右边界的不确定性。在汽车语料中，在“底板上只有一根引出线接在点火线圈上，点火线圈中心有磁性棒，高压点火线拧紧在点火线圈的木螺钉上。”中，“点火线圈”出现了3次，它的左邻接字有“在”和“，”，右邻接字有“上”、“中”和“的”。在整个语料中，字符串“点火线圈”总共出现了30次，不同的左邻接字有19个，右邻接字有21个，可见“点火线圈”的左右搭配词都很不固定，因此“点火线圈”很有可能是一个完整的词，进而可能是汽车术语。而在考察“点火线”是否完整时，发现“点火线”在整个语料中出现了33次，其不同的左邻接字有21个，右邻接字只有3个，则“点火线”不适合作为一个完整的词。

其中，左、右信息熵的公式如下所示:



其中，s是候选字符串，ls是s的左邻接字l和s结合所构成的字符串，p(ls|s)表示语料中出现s的情况下，s的左邻接字是l的条件概率。sr是s和s的右邻接字r结合构成的字符串，p(sr|s)表示语料中出现s的前提下，s的右邻接字为r的条件概率。LE(S)为字符串s的左信息熵，RE(S)为字符串s的右信息熵。LE(S)和RE(S)越大，说明左右邻接字越不固定，则s独立成词的可能性越大。为了综合评价S独立成词的可能性，通过给左、右信息熵设定相同的阈值来过滤不能独立成词的词。即如公式所示：



其中，为人工设定的阈值。

**③词频分布和信息熵相结合的方法**

在基于信息熵的术语抽取方法中，分别为左、右信息熵设定阈值存在局限性。不仅

在阈值调节时会比较繁琐，而且也不能很好地处理字符串的左、右信息熵的不平衡。在结合左、右信息熵的同时，综合考虑左右信息熵的不平衡性，采用如下信息熵公式：



在基于词频分布的术语抽取方法中，词频分布只考虑了术语的领域相关性，而没有考虑术语是否独立成词，因此很可能会抽取出一些结构不完整的字符串。词频分布公式中引入了平均词频，但用信息熵代替平均词频能很好的对术语是否独立成词加以判断，由此，得到如下公式：



**4、规则过滤**

使用的规则主要有词性规则、普通词过滤、边界判定是否合理、是否在其他语料中出现次数较多等。只考虑词频大于2的候选术语，且候选术语的长度小于10。使用的词性规则主要是术语词性的组合方式。

所采用的规则说明如下:

（1）词性规则主要有：

①术语中不能包含叹词、代词、处所词、状态词；

②术语不能以助词、连词、后缀开头；

③术语不能以前缀、方位词、连词、助词结尾；

④术语中必须含有名词、动词或量词成分。

（2）由于绝大多数的普通词都不是术语，本文用分词字典构建了一个普通词列表，用于过滤候选术语中的普通词，例如“上课”、“下雨”、“春天”等。

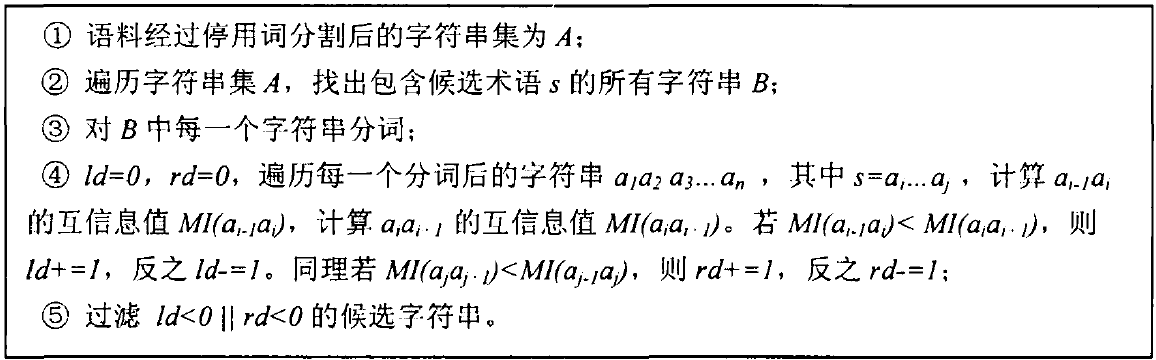
（3）虽然应用了信息熵，但还是有些候选术语不能独立成词。通过进一步的边界判定进行过滤，算法过程如图所示。

图2.3 边界判定算法

（4）该部分为后续研究，目前暂时不考虑。应用辅助的计算机语料，若候选术语在计算机语料中的词频超过在电动汽车语料中的词频的一半，则将候选术语过滤，当然这是在同样规模的语料的情况下。