阅读总结：

1、由于目前写文章进度实在是很慢，很多内容都不知道如何去清晰的表达，自己的方法也很难用文字来表示清楚，所以在阅读这篇文章的时候特别注意了一下别人在写文章时候是如何充分、完整、具体的表达自己的一些观点、方法、细节。

2、方法中所述的概率统计技术就是二元语法；规则方法就是过滤规则；和咱们处理标签搭配的方法十分接近，虽然这篇文章是的目的新词发现，但是表达方式还是值得我借鉴的。

论文题目：基于概率统计技术和规则方法的新词发现

作者：贾自艳，史忠檀

作者单位：中国科学院计算技术研究所

期刊名称及日期：计算机工程 2004 年 10 月

摘要：

该文分析了已有的短语抽取技术，并结合汉语特点，提出了基于概率统计技术和规则方法相结合的概念抽取方法。

该方法包括高效的“二元语法”统计模型、统计算法、统计选择策略、丰富的规则知识和规则过滤算法。

该方法适用于从大规模语料库中自动高效发现新词/短语。

1 概述

阐明新词发现所属领域；阐述新词发现的作用；列举了利用二元语法的文献；

对比了统计方法与规则方法；

对比了新词发现领域的统计方法与规则方法；

本文的工作。

本文在分析前人研究结果基础上，本着从实用角度出发，研究二者的融合方法实现新概念的抽取：以快速的统计方法为工具，自动获取特定领域的新词语、新概念；在此基础上通过一系列的规则进行过滤。这样既吸收统计方法的快速，又可保留专家系统方法的质量。

2 系统结构

基于概率统计和规则方法的新词发现系统结合了两种方法的有点，能够快速且高效的在大量的文本中发现高质量的概念。（叙述了流程图的工作流程，对一些模块的作用做了简要说明）

列举系统各个模块。

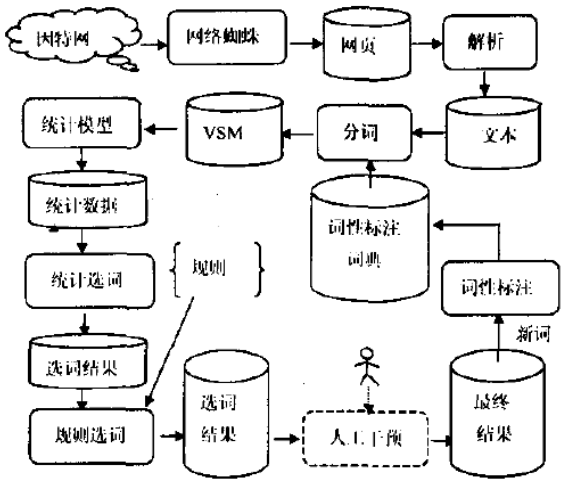

图1 概念发现系统的体系结构

3 统计模型

简单介绍 N 元文法并指出二元文法能够满足需求。

对模型进行定义。

定义1 “二元语法”的统计模型设计如下：统计的基本元素是系统所用的通用分词词典中的词。统计的对象是限于任何两个标点符号之间的连续词序列，可以表示为“$w_1\ w_2\ w_3\cdots w_n$”，从序列的第一个词语开始，依次记录相邻两个词语组合“$w_j\ w_{j+1}$”($1 \leqslant j \leqslant n-1$)的共现串、共现文档名称、共现位置等信息，同时实现共现次数、共现文档数目的累计。这样就得到两词集合(bi-word)，记为 $\text{bi-word} = \{w_j\ w_{j+1} | 1 \leqslant j \leqslant n-1\}$。为了便于描述，我们称第一个词“$w_j$”为首词，记为 Fword；第二个词“$w_{j+1}$”为尾词，记为 LWord。

3.1 数据预处理

HTML 网页-》文本-》分词

3.2 统计算法

为每个词生成一个倒排文件。然后统计每个词的出现次数、出现文档数目、与其他词的共现次数。

3.3 统计选词

Wiwj 组成新词的可信度定义为 wi 后出现 wj 的概率：p(wj|wi) = dfij/dfi。dfij 表示 wi 和 wj 共现次数，dfi 表示 wi 出现次数。针对每个 wi 分别计算与其共现所有 wj（设有 k 个）的

$$E(df_i) = \sum_{j=1}^{K} p(w_j \mid w_i) \times df_{ij}$$

共现频次均值：，这里定义统计选词原则：共现频次在均值之上的词汇组合是好的，即 dfij>E(dfi)。

对不好的组合，分析原因：

但不好的也不少，效率不是十分高。究其因如下：(1)统计公式的制定是把所有的词语都统一对待，只考虑它们的共性，没有考虑它们在使用过程中的种种个性，而这些个性往往是它们形成新词的决定性条件。(2)语料库本身的选取不能很好地满足上面理想化公式的使用条件。(3)汉语自身的灵活性。

引出规则方法的必要性：

总之，单纯的统计方法对于语料库的选择和统计公式的制订有极大的依赖性，难以达到很高的准确度。必须增加必要的知识，通过规则的方法来提高准确度。

4 规则选词

加入规则的好处。加入规则的原因。

4.1 单字组合词规则

不可组词；

表1 不可组词的单字词性规则

词性	标记	例子
数词	U***	“一”、“二”、“1”、“千”
代词	P***	“他”、“它”、“你”、“我”
介词	R***	“在”、“于”、“从”
助词	Z***	“的”、“么”、“哉”、“啊”
象声词	S***	“叭”、“嘿”、“砰”
姓氏单字	NSUR	“王”、“李”、“赵”、“贾”、“涂”

组合禁用词；

表2 单字组合词禁用词规则

规则	例子	数量
不可扩展的单字	“沪”、“斯”等以及二级字库中的偏僻字	500
只做首词的单字	“上”“前”等	14
只做尾词的单字	“除”、“加”、“至”等。量词（QOTH）和连词（COTH）	60

4.2 多字组合词规则

表3 多字组合词禁用词规则

规则	特点	组成/例子	数量
禁用虚词	没有组词能力，只有构造句子能力，同时自身是没什么实际意义词	助词、连词、疑问词、感叹词等虚词	1 047
禁用实词	自身意义相当完整，几乎没有必要再组词的词	成语、俗语、叠词，多为形容词、副词、动词等	5 675
只做首词	可用来扩展新词，但通常只做首词	“最高”、“依次”、“中和”等	77
只做尾词	可用来扩展新词，但通常只做尾词	“专业”、“折扣”、“障碍”等	414

多字组合词规律：

表4 多字词组合规律

规则	成词率	规则	成词率
单字+多字	低	名词+副词	低
名词+名词	较高	职位/职称+姓氏	低
名词+动词	低	前缀名词+词	高
名词+形容词	低	词+后缀名词	高

5 实验结果

语料：计算机世界语料库（40mb）、《人民日报》财经新闻语料库（20mb）

结果：

表5 新词发现结果示例

领域	计算机领域		经济领域	
类型	单字组合词	多字组合词	单字组合词	多字组合词
数量	3 000	3 010	1 586	3 550
例子	“死机”、“按钮”、“字节”、“字段”、“总机”、“声卡”、“站点”、“页眉”、“鼠标”、“网吧”、“黄页”、“师姐”等	“即插即用”、“北大方正”、“批处理”、“结束符”、“硬拷贝”、“局域网”、“差错 率”、“制造业”、“神经计算模型”、“演示版”、“计算机网络”等	“下岗”、“树叶”、“影碟”、“造假”、“韩国”、“招标”、“中标”、“转账”、“征婚”、“诊所”、“执肴”、“庄家”等	“安全感”、“八达岭长城”、“领导班子”、“包装箱”、“保健品”、“保障体制”、“项目经理”、“个体户”、“菜篮子”、“侏罗纪公园”等

新词发现结果

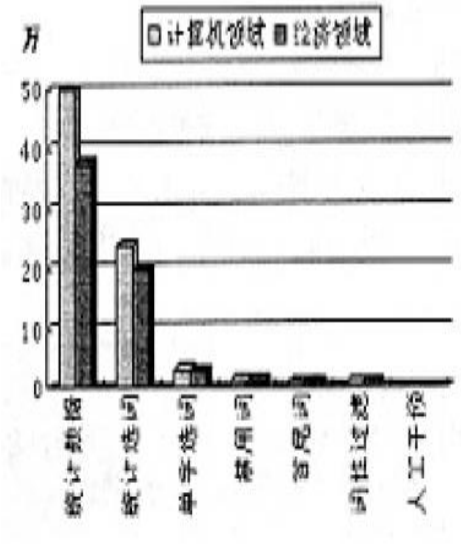

图2 单字词过滤过程结果显示

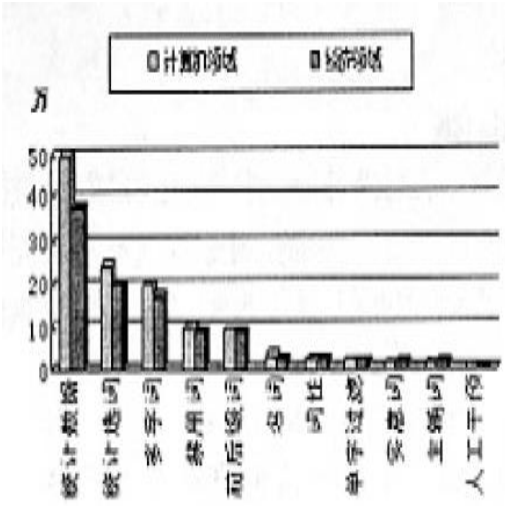

图3 多字词过滤过程结果显示

过滤结果

6 总结：简述方法及优势；指出方法不足以及下一步工作。