**专家信息抽取**

（一）国内外现状

信息抽取[1]的定义为从自然语言文本中抽取指定类型的指标、关系、事件等事实信息，并形成结构化数据输出的文本处理技术。

抽取专家信息包括专家的评价指标（人物的主要属性、论文发表情况、学术关系、专利情况、研究兴趣关键词）、描述指标（论文的统计数据、论文的被引用情况、论文的影响因子）和关联指标（论文合作关系、研究兴趣相似关系、共同参与课题的关系、同事关系、同学关系、人物与机构的关系）。

目前主流的抽取方法是给定某一类别的指标实例，从网页中抽取同一类别其他指标实例。基本思路在于种子词与目标词在网页中具有相同或者类似的上下文（包括网页结构和上下文）。因此需要首先利用种子词提取模板，随后利用模板提取更多同类指标。处理该问题的主流框架为：

指标抽取主要包括两种方法：一种是基于规则的抽取，一种是统计的抽取。

开始

种子

抽取模板

候选

打分器

结果

基于规则的指标抽取方案[2]依赖于规则模板，需要在人工总结现有的信息规则的基础上制定模板，优点是简单有效，缺点是非常依赖经验，可扩展性差。

基于统计的指标抽取方案都是以某些机器学习技术为核心，比如分类器、条件随机场、频繁模式挖掘和关联分析等。

张巧[3]等提出一种基于弱监督学习的属性抽取技术，在个人主页中，部分人物指标的相邻位置存在一些触发词，用以描述该指标，在bootstrapping方法中，这样的词被称为前导词。可以通过少量的前导词，即种子模式，根据已有信息，持续迭代更新相关指标特征以及模式，以此来发掘更多的模式特征。

贾真[4]等采用了一种基于弱监督学习的属性抽取技术，分类器采用的是最大熵模型。他在传统的n-gram模型上，改进成为n-pattern模型，两者区别在于，n-pattern中的项不一定连续，缓解了部分数据稀疏的问题。

（二）研究意义

从网络上收集给定领域的主要专家相关资料数据，并自动识别专家的各项指标，将整理好的指标以结构化的方式存储。

1. 研究内容

本文采用以条件随机场为核心的机器学习方法，对用网络爬虫爬取下来的特定文本数据，设计出一个高效的自动指标抽取工具。

研究内容主要以下几个部分：

1. 获取源语料并清洗

利用网络爬虫自动获取相关人物的个人简历，相关信息等。对粗糙的语料数据，进行分词并标注词性，并去除无关噪音。

1. 利用条件随机场训练抽取模型

把从非结构化文本中抽取专家人物指标这一问题，当作序列标注任务来求解，即实际上是使模型学习对特定人物指标的触发词和指标属性边界的相关特征。本文将使用条件随机场（CRF，Conditional Random Field Algorithm）技术来训练模型，并自动抽取专家人物指标。

1. 存储抽取的专家人物指标

将抽取好的专家人物指标以结构化的方式存储。本文将利用MySQL数据库来存储结构化的指标数据，利于后续研究时查阅和使用。

参考文献

1. 张丙奇, 姜吉发. 企业相关信息抽取技术研究与系统实现[J]. 微电子学与计算机, 2004, 21(1):1-6.
2. 丁君军, 郑彦宁, 化柏林. 国内外属性抽取研究综述[J]. 情报科学, 2011(5):793-796.
3. 张巧, 熊锦华, 程学旗. 基于弱监督学习的主页人物属性抽取方法[J]. 山西大学学报(自然科学版), 2015, 38(1).
4. 贾真, 杨燕, 何大可. 基于弱监督学习的中文百科数据属性抽取[J]. 电子科技大学学报, 2014, 43(5).