# 信息抽取研究与发展综述

/ 周玉新

【摘要】信息抽取的目的是旨在从海量数据中准确、快速地抽取出人们感兴趣的事实信息、并将其以结构化形式储存起来、以便于以后的分析和处理。文章回顾了信息抽取的历史、总结了信息抽取技术所使用的关键技术、并对其以后的发展提出了展望。

【关键词】信息抽取；自然语言处理；命名实体识别

【作者简介】周玉新，内蒙古民族大学计算机科学与技术学院。

## 1. 引言

随着计算机技术的飞速发展和互联网应用的普及，人们通过网络获取的资源呈现爆炸式增长，如何从不断增长的海量数据中快速、准确的抽取所需信息，已经成为研究人员所面临的一大重要问题，信息抽取（Information Extraction，IE）正是在这一背景下产生的。IE 的主要功能是从结构化、半结构化或非结构化文本中抽取出人们所需的特定事实信息，往往以结构化的形式描述这些被抽取出来的信息，并将其保存到数据库中，方便用户后续的查询和使用。[1]

最初的 IE 研究始于上个世纪六十年代中期，但是直到二十世纪八十年代末期，关于 IE 系统的研究才得到越来越多研究人员的关注，而这主要受益于从 1987 年召开的消息理解系列会议（MUC）。从 1987 到 1998 的 11 年间，MUC 共举行了七次会议，从第一届会议的探索性、没有任何明确任务定义到最后一次会议提出具体任务并制定严密的评测体系，包括场景模板填充、命名实体识别、共指消解等任务，MUC 确立并推动了信息抽取技术的研究与发展方向。

MUC 主要根据两个指标来评价 IE 系统：准确率和召回率。准确率是系统正确抽取的信息数与所有抽取信息数的比值；而召回率则是系统正确抽取的信息数与所有正确信息数的比值。在实际应用中还引入了另一个评价指标——F 值，它是准确率和召回率的加权平均，它可以对系统的性能进行综合评价，它的计算公式如下：

F- 值 =((1+α2)* 召回率 * 准确率 )/(( α2* 准确率 )+ 召回率 )

这里的 α 表示准确率和召回率的相对权重，最常用的值为 1，此时 F- 值即 F1 值。

继 MUC 会议之后，由 NIST 组织的 ACE 评测会议成为另一个推动信息抽取技术研究发展的主要动力。不同于 MUC，ACE 不再针对某个特定的领域或场景，而是采用基于漏报和误报为基础的另一套不同的评价体系，并且对系统的跨文档处理能力进行评测。这一新的会议将把 IE 技术的研究推向了新的高度。

## 2. 信息抽取研究中所使用的关键技术

在信息抽取研究中，主要利用相关的机器学习方法来自动抽取自由文本中的某些事实信息。所使用的具体机器学习技术主要分为基于统计和规则的方法。目前，信息抽取主要包括以下命名实体识别、指代消解、关系抽取以及事件抽取等几个方面的研究。[2]

2.1 命名实体识别。命名实体识别是信息抽取系统中最基本的工作，它的任务是识别出文本中的专有名词和有意义的数量短语，并进行实体分类和映射，以便于信息抽取的其他后续处理。

命名实体识别方法大体上可以分为三类：基于词典、基于规则和基于机器学习。基于词典的方法利用现有的数据资源定位文本中出现的命名实体。但是由于可用数据资源的匮乏以及新命名实体的不断出现，通常基于词典的方法所取得的精度和召回率都比较低。基于规则的方法则通过制定与各种特征有关的规则识别文本中的命名实体，但所制定的规则通常是非常具体的，有非常强的领域相关性，系统的可移植性非常差。目前，最常用的命名实体识别技术主要采用基于机器学习的方法，该方法使用训练数据学习对命名实体识别有用的特征，并利用学习到的特征识别文本中的命名实体。

2.2 指代消解。指代是广泛存在于自然语言各种表达中的一种常见语言现象，主要刻画文本概念之间的相互关联性，通常分为回指和共指两种类型。回指是指当前照应语与上文中出现的词、短语或句子之间存在密切的语义关联性；而共指则是指两个或两个以上的名词或名词短语指向同一参照体，回指不依赖于具体的上下文[3]。指代消解实际上是建立概念之间关联的过程，是文本处理的核心问题之一，MUC 早在第六届会议中就将其列为评测的子任务之一，在 IE 中起着重要作用。

指代消解方法主要分为两种：基于句法的方法和基于语料库的方法。基于句法的方法是早期进行指代消解时所采用的方法，其充分利用句法知识，并引入启发式学习方法来进行指代消解；而随着语言学的发展，出现了基于语料库的方法，其中主要有基于统计的方法和基于统计机器学习的方法等。[3]

2.3 关系抽取。关系抽取的作用是获取实体间的语法或语义关系，是 IE 的重要研究课题之一。在日常应用中，识别文本中的实体是信息抽取的第一步，更主要的是确定实体间的关系。与命名实体识别类似，实体关系的类型也是预先进行定义的，例如城市间的地理关系、人物和组织关系等。通常，人们将关系抽取问题看成是一个分类问题，最初是使用基于知识库的方法进行关系的抽取。但这种方法需要利用人工的方式构筑大规模的知识库，除了需要具有专业知识的专家之外，还需要付出繁重的重复劳动。

2.4 事件抽取。在信息抽取中，事件是指在某个特定的时间段及地域范围内所发生的，由一个或多个角色参与，由一个或多个动作所组成的一件事情。事件抽取主要研究如何从非结构化文本中抽取出用户所感兴趣的事件，同时用结构化的文本形式来描述这些事件，以供用户进行进一步的查询以及追踪分析等。事件抽取是 NLP 领域一个非常重要的研究方向，一直是事件抽取研究领域的关键问题。

## 3. 信息抽取的研究趋势

信息抽取经过近三十年的发展，已经成为 NLP 领域的一种重要分支，尤其是近年来网络技术的飞速发展极大地推动了信息抽取的发展。随着大数据时代的到来，传统的信息抽取方法可能无法适应大数据时代的处理方式；另外，系统性能和可移植性一直是制约信息抽取技术广泛应用的两大因素，如何克服和解决这两个问题，将是以后研究的一个重要方面。

参考文献：

[1] 郭喜跃，何婷婷 . 信息抽取研究综述 [J]. 计算机科学 ,2015(2):14-17.

[2] 张素香 . 信息抽取中关键技术的研究 [D]. 北京 : 北京邮电大学 ,2007.

[3] 孔芳，周国栋 . 指代消解综述 [J]. 计算机工程 ,2010(8):33-36.