<<中文信息抽取原理及应用>>

图书基本信息

书名：<<中文信息抽取原理及应用>>

13位ISBN编号：9787030266231

10位ISBN编号：7030266234

出版时间：2010-2

出版时间：科学出版社

作者：程显毅

页数：302

前言

随着计算机在各个领域的广泛普及和Internet的迅速发展，社会的信息总量呈指数级增长。
信息总量的量级，从20世纪90年代初的MB（10）过渡到GB（10）再到现在的TB（10）。
进入21世纪后，全世界信息总量更是以每三年增加一倍的速度递增。
据统计，在这些海量信息中，有60%－%是以电子文档的形式存在的。
为了应对信息爆炸带来的挑战，迫切需要一些自动化的技术帮助人们在海量信息中迅速找到自己真正需要的信息，信息抽取（informationex-traction，IE）正是解决这个问题的一种方法。
目前，对海量数据的操作主要还停留在信息检索阶段，即使是信息检索这个比较初级的任务，效果也很不理想：TREC2004TerabyteTrack的测试结果显示，文本信息检索的最高精度不超过30%。
扭转这种局面的出路在于IE的研究成果。
IE的任务是把无结构信息转换成有结构信息。
然而，限于目前的技术水平，印欧语言在I正方面的研究已经取得了一定的成果，但是中文IE研究相对滞后。
全书分两篇（原理篇共11章，应用篇共7章）。
原理篇主要讨论以下问题：（1）基于自然语言处理方式的信息抽取。
利用子句结构、短语和子句间的关系建立基于语法和语义的抽取规则实现信息抽取。
（2）基于规则的信息抽取。
由于规则较为集中地体现了领域知识和语言知识的融合，所以其构建过程即为知识的获取过程。
（3）基于统计模型的信息抽取。
基于规则的信息抽取是一种确定性的信息抽取模型，但并不是所有的自然语言现象都可以用确定性的规则来刻画的，而且这种规则的使用也具有不确定性。
在这种情况下，基于目前的语言学理论水平和计算技术条件，人们自然地会转向统计学方法，希望用在语料库中对相关数据的统计的方法，来描述自然语言的统计属性。
（4）基于认知模型的信息抽取。
语言理解具有明显的认知过程，因此，认知科学势必会对信息抽取产生积极的影响。
（5）命名实体识别、共指消解、模板填充、Web信息抽取等也是MUC规定的信息抽取任务。

<<中文信息抽取原理及应用>>

内容概要

由于网上的信息载体主要是文本，所以信息抽取技术对于那些把互联网当成是知识来源的人来说是至关重要的。
信息抽取系统可以看成是把信息从不同文档中转换成结构化数据系统。
因此，成功的信息抽取系统将把互联网变成巨大的数据库。
信息抽取技术是近十年来发展起来的新领域，遇到许多新的机遇和挑战。
　全书分两篇（原理篇共11章、应用篇共7章）。
原理篇主要讨论了信息抽取（IE）概念、任务、挑战和评测方法；基于NLP、统计、认知的信息抽取方法；命名实体识别、共指消解、模板填充、Web信息抽取等。
应用篇介绍了两个开发工具（GATE和WHISK），分析了IE在人机接口、电子交易、智能交通、竞争情报、问答系统、自动文摘等领域的应用。
　本书可作为本科高年级数据挖掘课程的参考书或研究生自然语言处理课程的教材，也可作为智能应用系统开发的参考资料。

<<中文信息抽取原理及应用>>

书籍目录

章节摘录

插图：随着计算机的普及以及互联网的迅速发展，大量的信息以电子文档的形式出现在人们面前。
信息的过量增长带来一定负面影响：面对巨量的信息，由于目前Web上存在的信息格式具有很大的异构性，信息之间的关联描述较少，用户通过直接浏览的方式获取所需的信息十分困难，用户不知道如何确切表达对真正想要的网上资源的需求（资源迷向），难以消化已经下载的信息（信息过载）。
如何将大量无序的信息及时准确地进行抽取、过滤、归类组织成便于查询检索的形式’已成为研究开发的焦点。
迫切需要一些自动化的工具帮助人们在海量信息源中迅速找到真正需要的信息，信息抽取（informationextraction，IE）研究正是在这种背景下产生的。
具体来讲就是：（1）互联网已经成为一个巨大的隐式信息源。
（2）垂直搜索发展迅速。
（3）传统信息检索（informationretrival，IR）方法已无法满足现代社会发展的需求。
（4）大量信息需要结构化。
（5）传统的基于HTML的抽取方法应用受限。
（6）中文自然语言处理技术的发展带来契机。
信息抽取的目标是把文本里包含的信息进行结构化处理，变成表格一样的组织形式。
信息抽取系统的输入是原始文本，输出的是固定格式的信息点。
信息点从各种各样的文档中被抽取出来，然后以统一的形式集成在一起。
信息以统一的形式集成在一起的好处是方便检索和比较，如比较不同的招聘和商品信息。
还有一个好处是能对数据进行自动化处理，如用数据挖掘方法发现和解释数据模型。
信息抽取技术并不试图全面理解整篇文档，而只是对文档中包含相关信息的部分进行分析。
至于哪些信息是相关的，那将由系统设计时定下的领域范围而定。
信息抽取技术对于从大量的文档中抽取需要的特定事实来说是非常有用的。
互联网上就存在着这么一个文档库。
在网上，同一主题的信息通常分散存放在不同网站上，表现的形式也各不相同。
若能将这些信息收集在一起，用结构化形式储存，那将是有益的。

编辑推荐

《中文信息抽取原理及应用》由科学出版社出版。

<<中文信息抽取原理及应用>>

版权说明