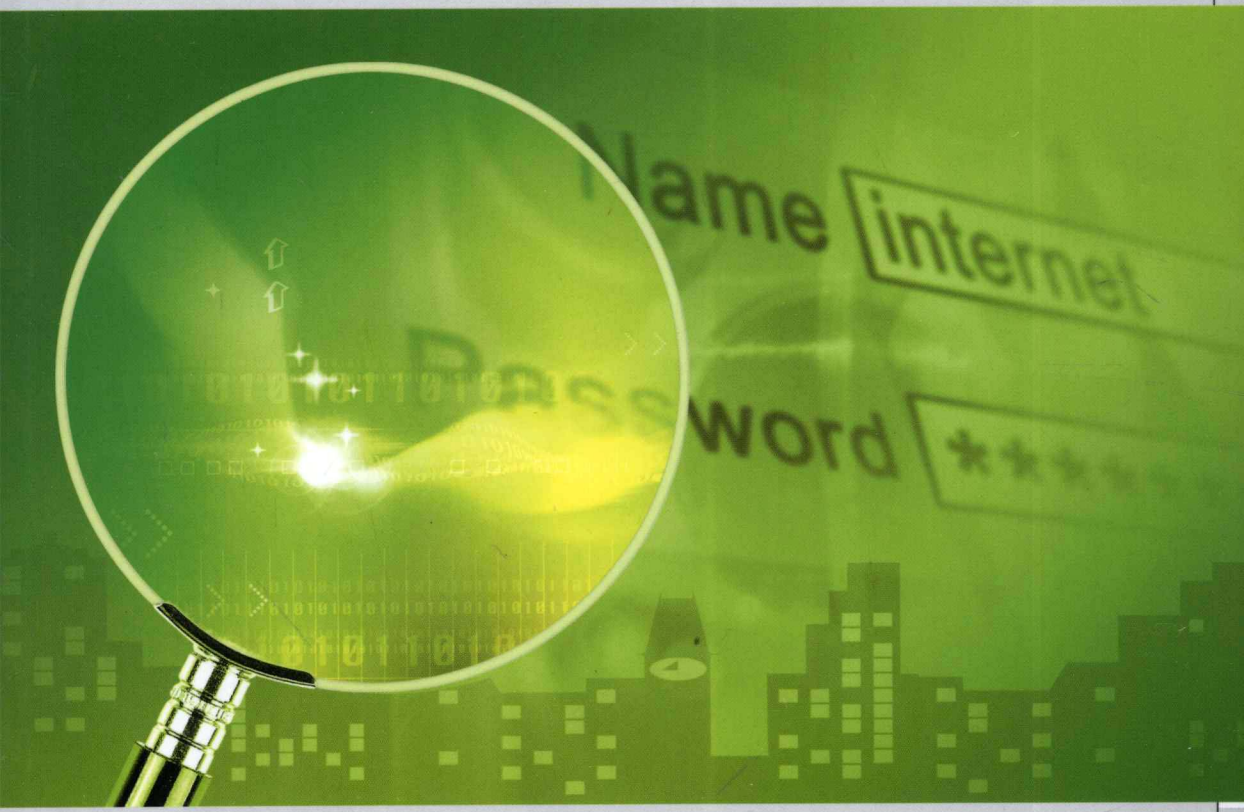
程显毅 朱 倩 王 进 编著

中文信息抽取原理及应用



科学出版社
www.sciencep.com

有关此电子书的说明

本人可以帮助你找到你要的PDF电子书，计算机类，文学，艺术，设计，医学，理学，经济，金融等等。质量都很清晰，为方便读者阅读观看，每本100%都带可跳转的书签索引和目录，只要您提供给我书的相关信息，一般我都能找到，如果您有需求，请联系我 QQ1779903665。

PDF代找说明：

本人已经帮助了上万人找到了他们需要的PDF，其实网上有很多PDF,大家如果在网上不到的话，可以联系我QQ，大部分我都可以找到，而且每本100%带书签索引目录。因PDF电子书都有版权，请不要随意传播，如果您有经济购买能力，请尽量购买正版。

提供各种书籍的pd电子版代找服务，如果你找不到自己想要的书的pdf电子版，我们可以帮您找到，如有需要，请联系 QQ 1779903665.

备用:QQ 461573687

声明：本人只提供代找服务，每本100%索引书签和目录，因寻找和后期制作pdf电子书有一定难度，仅收取代找费用。如因PDF产生的版权纠纷，与本人无关，我们仅仅只是帮助你寻找到你要的pdf而已。

(TP-4577.0101)

中文信息抽取原理及应用

ISBN 978-7-03-026623-1

9 787030 266231 >

科学出版社

电　话：010-64000249

E-mail：gcjs@mail.sciencep.com

销售分类建议：计算机

定　价：58.00元

中文信息抽取原理及应用

程显毅　朱　倩　王　进　编著

科学出版社

北京

内 容 简 介

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。

图书在版编目(CIP)数据

中文信息抽取原理及应用/程显毅，朱倩，王进编著. —北京：科学出版社，2010.2

ISBN 978-7-03-026623-1

Ⅰ.中… Ⅱ.①程…②朱…③王… Ⅲ.汉语-文字处理系统-研究 Ⅳ.TP391.1

中国版本图书馆 CIP 数据核字（2010）第 019411 号

责任编辑：张海娜/责任校对：刘小梅

责任印制：赵 博/封面设计：嘉华永盛

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2010 年 2 月第 一 版 开本：B5（720×1000）

2010 年 2 月第一次印刷 印张：19 3/4

印数：1—3 000 字数：383 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

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（10^6）过渡到 GB（10^9）再到现在的 TB（10^{12}）。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%～70%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（information extraction，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：TREC 2004 Terabyte Track 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 IE 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。

应用篇首先介绍了两个开发工具，即 GATE 和 WHISK；然后讨论 IE 在自然语言查询接口、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。

程显毅老师编写了第 1、3～10、18 章；王进老师编写了第 2、11～13 章；朱倩老师编写了第 14～17 章。最后由程显毅老师统稿。

感谢孙萍、史燕、杨天明、蔡月红、陈海光等在资料整理过程中所做的工作。

感谢南通大学计算机学院、江苏大学计算机学院给予的支持。

本书得到国家自然基金（60873069）、江苏省研究生创新计划项目（CX09B-2042）、南通大学自然科学类科研基金（092023）的资助。

信息抽取领域的研究发展迅速，对许多问题作者并未作深入研究，一些有价值的新内容也来不及收入本书，加上作者知识水平和实践经验有限，书中难免存在不足之处，敬请读者批评指正。

目　录

前言

原　理　篇

第 1 章　绪论 …… 3
1.1　信息抽取产生的背景 …… 3
1.2　信息抽取概念 …… 4
1.3　信息抽取任务 …… 5
1.4　信息抽取和相关概念之间的关系 …… 6
1.5　信息抽取的意义 …… 10
1.6　信息抽取的研究现状 …… 12
1.6.1　国外研究现状 …… 12
1.6.2　国内研究现状 …… 14
1.7　存在的问题及解决策略 …… 15
1.8　信息抽取的挑战和趋势 …… 16
第 2 章　信息抽取评估 …… 19
2.1　信息抽取评估一般原则 …… 19
2.2　国际测评会议 …… 20
2.2.1　MUC 测评会议 …… 21
2.2.2　ACE 测评会议 …… 21
2.2.3　MET 测评会议 …… 26
2.2.4　DUC 测评会议 …… 27
第 3 章　信息抽取原理 …… 28
3.1　信息抽取系统体系结构 …… 28
3.2　信息抽取方法分类 …… 30
3.3　文本表示 …… 31
3.3.1　向量空间模型 …… 31
3.3.2　*N*-gram 模型 …… 33
3.3.3　类短语串模型 …… 33
3.3.4　概念模型 …… 37
3.3.5　事件模型 …… 39
3.3.6　图模型 …… 40
3.4　词法分析 …… 41

3.4.1 自动分词 …… 41
3.4.2 词性标注 …… 44
3.5 语义标注及其角色 …… 45
3.5.1 语义标注 …… 45
3.5.2 语义角色精细等级 …… 47
3.5.3 框架网及其语义角色 …… 49
3.5.4 命题库及其语义角色 …… 52
3.5.5 中文网库及其语义角色 …… 56
3.5.6 问句问点的语义角色 …… 60
3.5.7 语义标注方法及步骤 …… 61
3.6 语料库建设 …… 62
3.6.1 语料库在信息抽取研究中的地位 …… 63
3.6.2 大型现代汉语语料库简介 …… 64
3.6.3 语料库系统 …… 66
3.6.4 语料库标注 …… 70
第 4 章 基于 NLP 的信息抽取 …… 71
4.1 经典系统 …… 71
4.2 相关技术 …… 72
第 5 章 基于规则的信息抽取 …… 77
5.1 原理 …… 77
5.2 规则的建立 …… 80
5.3 规则抽取系统 …… 84
5.4 自由文本规则抽取系统讨论 …… 89
5.5 规则抽取系统比较 …… 91
5.6 规则抽取的困难 …… 92
第 6 章 基于统计模型的信息抽取 …… 94
6.1 原理 …… 94
6.2 *N* 元模型 …… 94
6.2.1 基本思想 …… 94
6.2.2 数据平滑方法 …… 95
6.3 基于隐马尔可夫模型的信息抽取 …… 96
6.3.1 马尔可夫模型 …… 96
6.3.2 隐马尔可夫模型 …… 97
6.3.3 隐马尔可夫模型的三个基本问题 …… 98
6.3.4 基于隐马尔可夫模型的信息抽取 …… 102
6.4 最大熵模型 …… 104
6.4.1 形式化描述 …… 104

6.4.2 模型求解 …… 105
6.5 条件随机场模型 …… 106
6.5.1 形式化描述 …… 106
6.5.2 参数估计 …… 107
6.5.3 特征选择 …… 108
6.6 支持向量机模型 …… 109
6.6.1 线性 SVM …… 110
6.6.2 线性 SVM 构造 …… 111
6.6.3 非线性 SVM …… 112
6.6.4 非线性 SVM 构造 …… 113
6.6.5 SVM 学习算法 …… 113
6.7 统计模型的局限性 …… 114
第 7 章 基于认知模型的信息抽取 …… 116
7.1 原理 …… 116
7.2 基于本体的信息抽取 …… 116
7.2.1 本体的概念 …… 116
7.2.2 本体建模 …… 117
7.2.3 本体描述 …… 119
7.2.4 基于本体的信息抽取逻辑结构 …… 121
7.2.5 应用实例 …… 123
7.3 基于知网的信息抽取 …… 126
7.3.1 引言 …… 126
7.3.2 义原 …… 128
7.3.3 概念表示 …… 130
7.3.4 基于知网的中文信息结构抽取研究 …… 132
7.4 基于 HNC 理论的信息抽取 …… 135
7.4.1 HNC 理论的研究目标和研究内容 …… 135
7.4.2 HNC 理论的语言概念空间 …… 138
7.4.3 HNC 理论的概念表述模式 …… 141
7.4.4 HNC 理论的语句表述模式 …… 143
7.4.5 语句相似度计算 …… 145
7.4.6 基于 HNC 的语境框架抽取 …… 146
7.5 基于混合模型的信息抽取 …… 150
第 8 章 中文命名实体识别 …… 151
8.1 命名实体 …… 151
8.2 中文人名识别 …… 152
8.2.1 中文姓名用字特点 …… 152

8.2.2 中文姓名前后文规律 …… 153
8.2.3 基于规则的识别模型 …… 153
8.2.4 基于统计的识别模型 …… 154
8.3 中文地名识别 …… 157
8.3.1 地名识别知识库的建造 …… 157
8.3.2 地名识别规则库建造 …… 162
8.3.3 地名识别推理机制 …… 163
8.3.4 地名自动识别系统的实现 …… 164
8.3.5 示例和实验结果 …… 168
8.4 中文机构名识别 …… 169
8.4.1 机构名特点 …… 169
8.4.2 模型概述 …… 170
8.4.3 标注体系 …… 171
8.4.4 后界判断 …… 172
8.4.5 前部标注 …… 175
8.4.6 机构名识别过程 …… 180
8.5 数量结构识别 …… 181
8.5.1 数量结构的类型及自动识别的意义 …… 181
8.5.2 程序的算法设计及总流程 …… 182
第 9 章 共指消解 …… 185
9.1 指代的解析 …… 185
9.2 歧义问题 …… 186
9.3 测评标准 …… 187
9.4 相关技术 …… 188
9.4.1 国外的相关技术 …… 188
9.4.2 国内的相关技术 …… 191
9.5 中文的共指消解 …… 193
第 10 章 信息抽取模板 …… 195
10.1 模板的定义和结构 …… 195
10.2 信息结构抽取 …… 195
10.3 事件探测 …… 196
10.4 模板生成 …… 196
10.4.1 模板元素（TE）的构建 …… 197
10.4.2 模板关系（TR）的构建 …… 198
10.4.3 场景模板（ST）的产生 …… 200
10.5 模板填充 …… 201

第 11 章　Web 信息抽取 …… 203
11.1　概述 …… 203
11.2　语义 Web …… 203
11.2.1　基本概念 …… 203
11.2.2　本体描述语言 …… 205
11.3　格式转换 …… 206
11.4　信息解析 …… 206
11.5　基于 DOM 子树的抽取规则抽取算法 …… 207
11.5.1　DOM …… 207
11.5.2　XPath …… 208
11.5.3　XSLT …… 210
11.5.4　NE-DOM 分析 …… 210
11.5.5　基于 DOM 子树的抽取规则抽取算法 …… 212

应　用　篇

第 12 章　信息抽取工具 GATE …… 219
12.1　概述 …… 219
12.1.1　GATE 的组件 …… 219
12.1.2　GATE 的作用 …… 221
12.1.3　GATE 的应用 …… 221
12.1.4　GATE 系统的整体架构 …… 225
12.2　英文信息抽取 …… 226
12.2.1　信息抽取插件 ANNIE …… 226
12.2.2　抽取规则插件 JAPE …… 226
12.2.3　GATE 中的标注集的数据结构分析 …… 228
12.2.4　批量的英文信息抽取 …… 229
12.3　中文信息抽取 …… 230
12.3.1　中文信息抽取的困难 …… 230
12.3.2　基于 GATE 的中文信息抽取系统的解决思路 …… 231
12.4　GATE 组件扩展 …… 233
第 13 章　信息抽取工具 WHISK …… 235
13.1　WHISK 的规则表示 …… 235
13.1.1　结构化和半结构化文本的规则 …… 235
13.1.2　语法文本的扩展规则 …… 236
13.2　WHISK 算法 …… 238
13.2.1　人工标记训练样本 …… 238
13.2.2　从种子例子中创建一条规则 …… 239
13.2.3　槽的抽取 …… 240

13.2.4 增加术语到建议的规则上 …… 242
13.2.5 爬山和地平线效应 …… 243
13.2.6 预删除和后删除的规则 …… 243
13.3 训练集合构造 …… 244
13.3.1 选择样本 …… 244
13.3.2 何时停止标注 …… 245
13.4 实验分析 …… 245
13.4.1 问题描述 …… 245
13.4.2 方法和指标 …… 247
13.4.3 实验及分析 …… 247
13.5 关于 WIHSK 的讨论 …… 252
第 14 章 IE 在自然语言查询接口中的应用 …… 254
14.1 自然语言查询接口的背景 …… 254
14.2 自然语言查询接口的逻辑结构 …… 254
14.3 信息抽取模型 …… 257
14.4 信息抽取算法 …… 258
第 15 章 IE 在国民经济中的应用 …… 260
15.1 面向电子交易的信息抽取模型 …… 260
15.1.1 总体框架 …… 260
15.1.2 基于 DOM 树的抽取规则 …… 262
15.2 城市道路交通的信息抽取 …… 265
15.2.1 城市道路交通信息抽取的技术内涵 …… 265
15.2.2 城市道路交通信息抽取技术框架 …… 267
15.3 IE 在竞争情报研究中的应用 …… 268
第 16 章 基于自然语言处理的研究主题抽取 …… 271
16.1 问题描述 …… 271
16.2 研究主题抽取 …… 273
16.3 多语环境下的关键词语抽取 …… 274
16.4 研究主题聚类 …… 276
16.5 研究主题分析的实验结果 …… 278
第 17 章 IE 在自动文摘中的应用 …… 285
17.1 问题描述 …… 285
17.2 单文档自动文摘 …… 285
17.2.1 自动文摘过程 …… 285
17.2.2 自动文摘方法 …… 286
17.3 多文档自动文摘 …… 288

17.4　自动文摘系统的测评 …… 291
第 18 章　IE 在问答系统中的应用 …… 294
18.1　概述 …… 294
18.1.1　研究背景 …… 294
18.1.2　问答系统分类 …… 295
18.1.3　研究现状 …… 295
18.2　问答系统关键问题研究 …… 297
18.2.1　问题分析 …… 297
18.2.2　问题理解 …… 297
18.2.3　信息检索 …… 299
18.2.4　答案抽取 …… 299
参考文献 …… 303
结束语 …… 304

原　理　篇

自然语言理解的研究是一项激动人心的工作，是下一代计算机革命的主要动力

試車票

第1章 绪 论

1.1 信息抽取产生的背景

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（information extraction，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（information retrival，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。

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做是把信息从不同文档中转换成数据库记录的系统。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。

1.2 信息抽取概念

信息抽取技术属于知识技术中知识发现的范畴，它的宗旨就是在文本中分析处理大量的数据，发现有用的知识，为用户提供所需问题的答案。

信息抽取的宗旨在于抽取指定的信息，它突破了信息检索中必须由人来阅读、理解、抽取信息的局限性，实现了信息的自动查找、理解和抽取。信息抽取的定义主要有以下几种。

定义 1.1 信息抽取是以一个以未知的自然语言文本作为输入，产生固定格式、无歧义的输出数据或事实的过程。

定义 1.2 信息抽取是将获取的信息根据预先定义的模板，从文本抽取特定的信息，形成结构化的数据，帮助人们对信息内容进行分析和整理。

定义 1.3 信息抽取是指从一个给定的文档集合中自动识别出预先设定的实体、关系和事件等类型信息，并对这些信息进行结构化存储和管理的过程。

定义 1.4 信息抽取的任务是从文本中抽取字符形式的信息，并将此信息填入带标记的槽中，来表明其含义。

定义 1.5 信息抽取以空槽的形式提出应从原文中获取的文摘框架各项内容。

定义 1.6 信息抽取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。输入信息抽取系统的是原始文本，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。

抽取后的数据和事实可以直接向用户显示，也可作为原文信息检索的索引，或存储到数据库、电子表格中，以便于以后的进一步分析。比如，从新闻报道中抽取出恐怖事件的详细情况：时间、地点、作案者、受害者、袭击目标、使用的武器等；从经济新闻中抽取出公司发布新产品的情况：公司名、产品名、发布时间、产品性能等；从病人的医疗记录中抽取出症状、诊断记录、检验结果、处方等。通常，被抽取出来的信息以结构化的形式描述，可以直接存入数据库中，供用户查询以及进一步分析利用。

信息抽取涉及两个方面的因素：①用户从待分析的文本集中指定感兴趣的信息（相似度计算）；②过滤文本集并以一定的格式输出匹配的信息（语义模板填充）。

一般来说，信息抽取系统的处理对象是自然语言文本，尤其是非结构化文本。但从广义上讲，除了电子文本以外，信息抽取系统的处理对象还可以是语音、图像、视频等其他媒体类型的数据。本书只讨论狭义上的信息抽取，即针对自然语言文本的信息抽取。

1.3 信息抽取任务

为了抽取指定的信息，一般而言，信息抽取系统需要完成下面的具体任务：

(1) 准确识别文本中的各种命名实体（named entity，NE），这一般包括文本中出现的人名、地名、机构名、时间、货币以及各种数字等。

(2) 准确识别并标注指称相同的不同语言元素，即共指（co-reference，CO）。例如在下面的例子中，用黑体标注的部分均指同一对象。

例 1.1 中央处理单元的英文缩写为 **CPU**，**它**由运算器和控制器组成。

(3) 利用领域知识进行推理，在实体-实体之间、实体-事件之间建立关系。例如，根据处理的文本建立人名和机构名之间的隶属关系，机构名和地名之间的关系，如“李梅是南通大学老师”、“中国银行投资的国家是中国”等。

消息理解会议（message understanding for comprehension，MUC）将信息抽取过程简化为如表 1.1 所示的五个（由简到难的）阶段。

表 1.1 信息抽取的五个阶段

名称	作用
命名实体识别（name entity，NE）	查找并且对名字、地点等进行分类
共指消解（co-reference，CO）	鉴别文本中的实体之间的恒等关系式
模板元素构建（template element，TE）	为命名实体识别结果添加描述信息（使用 CO）
模板关系构建（template relation，TR）	在 TE 实体之间找出关系
情景模板建立（scenario template，ST）	把 TE 和 TR 的结果放到合适的具体事件情景下

NE、CO、TE、TR、ST 需要对文本进行“适度的”（浅层、非完整的）词法、句法及语义分析，并作各种标注。需要使用合适的词典、构词规则库等知识库的支持。使用模式匹配方法识别指定的信息（即找出信息模式的各个部分）。

MUC 详细地给定了每次测评内容的各个任务，近年来 MUC 的测试结果是令人鼓舞的，它表明现有的许多系统已经具备了相当程度的大规模真实文本的处理能力。具体任务是：

(1) 对网页进行解析以及相关处理。

(2) 对文本进行中文分词和实体识别。

(3) 对包含事实信息的文本块进行抽取。

(4) 对过滤文本使用上下文规则分析。

(5) 用全局最优的关系和事件填充模板。

(6) 相关知识库的构建：分词与实体识别词典、领域特征词典、文本分块规则、上下文分析规则、预定义信息模板等。

1.4 信息抽取和相关概念之间的关系

1. 信息抽取和信息检索之间的关系

与信息抽取密切相关的一项研究是信息检索，但信息抽取与信息检索存在差异，主要表现在以下三个方面。

(1) 功能不同。信息检索系统主要是从大量的文档集合中找到与用户需求相关的文档列表；信息抽取系统则旨在从一个文本中直接获得用户感兴趣的事实信息。

(2) 处理技术不同。信息检索系统通常利用统计及关键词匹配等技术，把文本看成词的集合，不需要对文本进行深入分析理解；信息抽取往往要借助自然语言处理技术，通过对文本中的句子以及篇章进行分析处理后才能完成。

(3) 适用领域不同。由于采用的技术不同，信息检索系统通常是领域无关的，而信息抽取系统则是领域相关的，只能抽取系统预先设定好的有限种类的事实信息。

信息抽取技术可视为信息检索技术的一个深化。信息检索从文档的集合中寻找与用户要求相关的文本或段落。信息抽取则是在相关文本或段落的基础上，发现用户需要的信息。这两种技术是互补的，信息抽取系统通常以信息检索系统的输出作为输入。例如，由信息检索系统寻找相关文档，而后由信息抽取系统在相关文档中抽取所需信息；反之，也可在信息抽取的基础上，进行高精度的信息检索，二者的结合能够更好地服务于用户的信息处理需求。

信息检索一般对文本的语义不进行分析，而由用户对文本的语义做出解释。信息抽取则由系统分析文本的语义在此基础上给出用户需要的信息。

2. 信息抽取和自然语言理解之间的关系

信息抽取是自然语言理解技术和实际应用相折中的产物。由于语言本身的复杂性以及人们对语言理解机制缺乏深刻认识，目前的自然语言处理水平尚不能对任意的文本进行深入分析，且不具备深入理解自然语言的能力；对自然语言文本的完全理解仍是遥远的梦想。信息抽取有时称为消息理解，这从美国国防高级研究计划局（Defence Advanced Research Projects Agency，DARPA）倡导并资助以交流和测评信息抽取技术的会议名称上可见一斑，这个会议被称为消息理解会议。然而需要指出的是，信息抽取中有关理解的含义和传统的自然语言理解的含义并不完全相同。首先，二者的目标不完全相同，信息抽取追求对文本的有限理

解，不是真正的文本理解。其次，在信息抽取中，用户一般只关心有限的感兴趣的事实信息，它的主要功能是根据预先设定的任务，抽取特定类型的信息，而不关心文本意义的细微差别以及作者的写作意图等深层理解问题。因此，信息抽取只能算是一种浅层的或者说简化的文本理解技术。再次，在信息抽取研究中，对“理解”的定义比较具体，具有可以操作和测评特点，理解意味着指定信息的成功抽取，而在传统的自然语言理解研究中，理解一词的含义并不明确，什么算是成功的理解，什么不算成功理解，缺乏可以操作的客观标准，不同的人会对一个“理解”的结果有不同的评价。最后，信息抽取系统在对一篇文本进行处理之前，清楚地知道理解的任务和深度，这些都可以通过预先设定的模板做了明确规定。

3. 信息抽取和问答系统（question answering，QA）之间的关系

QA 是指从一个有限大的文档集合 $D=\{d_1,d_2,\cdots,d_n\}$ 中，针对用户用自然语言语句所提出的问题 Q，找出简短精确的答案 A 的过程。

QA 一般要经过三个处理阶段：问题分析阶段、相关文档查找阶段和答案抽取与排序阶段。问题分析是指对问题进行分析，判断出问题所属的问答类型，并分离或推导出据以进行文档查找的关键词；相关文档查找是指根据问题分析阶段得到的关键词从源文档集合中查找相关文档；答案抽取与排序是指从得到的相关文档中进行候选答案的抽取并按照某种标准对抽取出的候选答案进行排序。

一个问答类型将一类问题所期待答案的语义类、该类问题所对应的问题模式和该类问题所对应的答案模式关联了起来。有了这么一个问答类型体系，则对于一个提问 Q，通过分析 Q 得出它所对应的问题模式，通过它所对应的问题模式找出它所属的问答类型，然后使用该问答类型所对应的答案模式从文本中查找问题 Q 的答案。因而，关键是建立一个领域无关的问答类型体系并找出与问答类型体系中每个问答类型相对应的答案模式。这就需要信息抽取中的模式获取技术的支持。所以，QA 是 IE 的一个用场，而 IE 的最终作用可以通过 QA 体现出来（见图 1.1）。

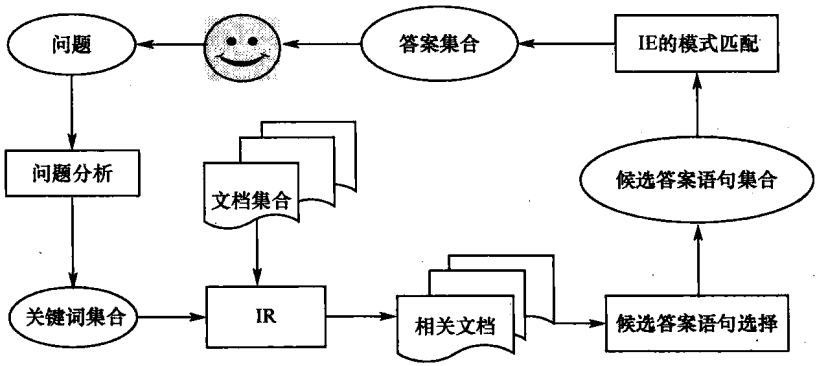


图 1.1 从 QA 的一般实现步骤看 QA 与 IE、IR 之间的关系

4. 信息抽取和文本挖掘之间的关系

信息抽取将非结构化的自然语言文本映射为相应的结构化数据。主流技术是基于机器学习的，以演绎推理为基础。文本挖掘从非结构化的自由自然语言文本集中发现隐含知识，要在掌握一定的文本“内容脉络”的基础上才能真正实现，以归纳推理为主。所以，文本挖掘的前期步骤间接地利用了信息抽取技术。

典型的文本挖掘过程如图 1.2 所示。首先从文本中抽取适当的特征，将文本表示成计算机能够理解的数字形式。根据处理速度和精度的需要，可以对文本中的特征进行选择优化。然后采用各种文本挖掘方法发现隐藏的知识模式，以满足用户评价标准的模式最终输出，成为指导人们实践的有用知识。

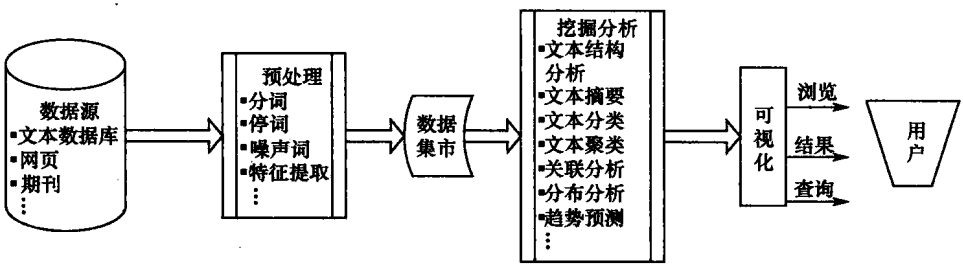


图 1.2 文本挖掘过程

5. 信息抽取和文本聚类之间的关系

文本聚类是文本组织、自动话题抽取、快速信息抽取和过滤的关键技术，文本聚类的实际应用包括：可以作为自然语言处理应用的预处理步骤，例如对多文档的总结；对搜索引擎返回结果重新聚类，可以帮助读者在更短的时间内找到它们所需的资料；可以通过对用户感兴趣的文章进行聚类，可以抽取出用户的阅读喜好模型；产生有效的文本分类器。

没有任何一种文本聚类技术（聚类算法）可以普遍适用于揭示各种多维数据集所呈现出来的多种多样的结构。根据数据在聚类中的积聚规则以及应用这些规则的方法，有多种聚类算法。

6. 信息抽取和信息过滤之间的关系

信息过滤本质上是一个两类分类问题，既可以用来将用户反感的信息滤掉，如黄色、淫秽、反动信息，也可以用来将用户感兴趣的信息过滤出来，主动地推送给用户，方便了用户快速准确地获得信息。它将综合运用分类技术和摘要技术，从应用角度看信息过滤是信息抽取的逆过程或主动的信息抽取过程。信息抽

取和相关概念关系如图 1.3 所示。

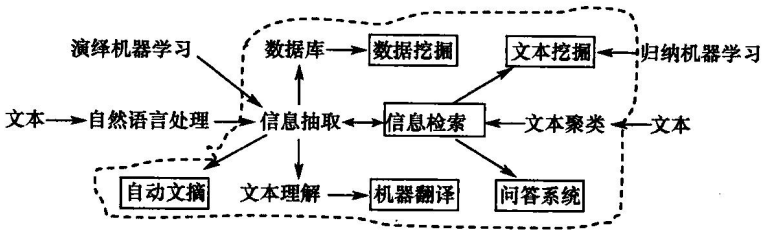


图 1.3 信息抽取和相关概念关系

图 1.3 中，虚线框起来的部分为文本处理（知识发现），方框部分为文本处理的输出。文本可以有两种途径作为文本处理的输入。从图 1.3 中可以看出，信息抽取是文本处理的核心，自然语言处理和机器学习是信息抽取的基础。基于信息抽取的文本挖掘技术是知识发现的研究趋势。

7. 信息抽取与知识发现之间的关系

从知识发现的发展过程来看，存在着两条清晰的研究路线：一条是基于数据挖掘（data mining，DM），其主要研究的是从结构化的数据（如数据库中的数据）中发现新的知识；另一条是从自然语言处理（NLP）出发，主要研究如何从非结构化或半结构化的数据（如 Word、HTML 或 PDF 文件）发现新知识。

信息抽取处于知识技术的初始阶段。图 1.4 是整个知识技术的两个阶段，信息抽取处于初始阶段，也就是图中左下角的“LAS 信息抽取系统”。经过信息抽取之后，无结构的数据被加工成为有结构的信息。

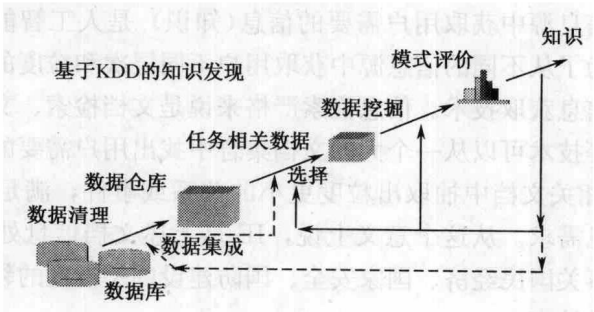


图 1.4 知识发现的各个阶段

图 1.5 中的第一阶段（左面方框内），知识系统获得海量的网络数据库资源、Internet 资源以及本地数据资源，并将这些资源储存到本地信息存储中，此时的信息属于无结构信息。之后，经过信息抽取过程，将这些无结构的数据转换成为有结构的信息，并且被储存到知识仓库中，这就是知识技术的第一阶段。

图 1.5 中的第二阶段（右面方框内），利用相应的组件对知识仓库中结构化的知识进行数据挖掘、决策支持或者问题解答等更深入的应用。

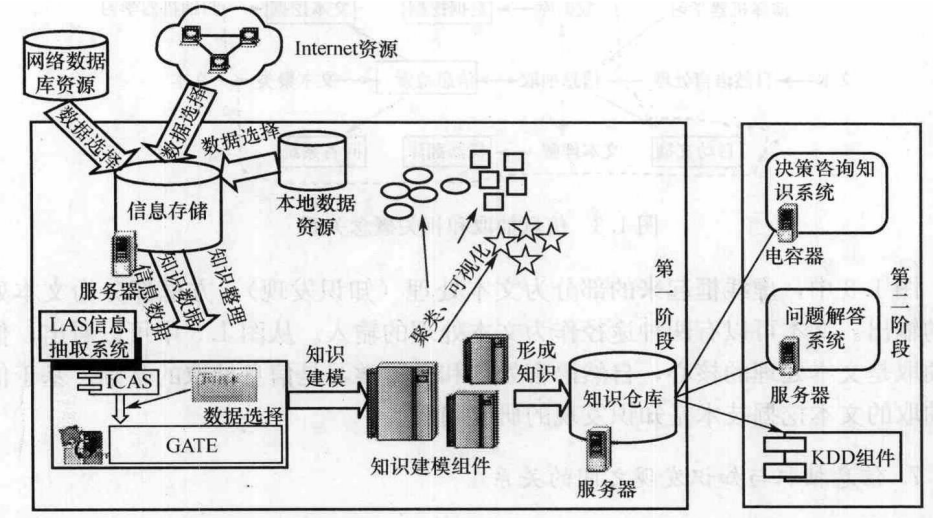


图 1.5 知识发现的整个过程以及信息抽取在其中的地位

1.5 信息抽取的意义

1. 从满足用户信息需求的角度来看，IE 是其他信息获取手段的一种有益补充

随着 Internet 的发展，网络上的文本文档数目巨大且正在快速增长。如何从如此巨大的网络信息源中获取用户需要的信息(知识）是人工智能和 Internet 研究的一个主题。为了从不同的信息源中获取用户不同层次和粒度的信息，人们发明了各种不同的信息获取技术。信息检索严格来说是文档检索、文本分类、文本过滤、文本聚类等技术可以从一个大的文档集合中找出用户需要的相关文档，而 IE 技术却可以从相关文档中抽取出粒度更小的关系或事件，满足用户更深层次和更细粒度的信息需求。从这个意义上说，IE 是上述文档信息处理技术的一种有益补充。许多事关国民经济、国家安全、国防建设的关键性的领域和广大的网络用户都需要这种补充。

2. 从技术实现的角度来看，IE 为其他信息获取技术提供支持

IE 作为一种将非格式化信息转换为格式化信息的一种手段，为进一步的信息处理如数据库查询、数据挖掘、文本挖掘等打下了基础。此外，还能对信息检索（IR)、知识问答（QA)、个性化信息服务等的实现起功能上的支持作用，或

者提高它们的性能（见图 1.6）。

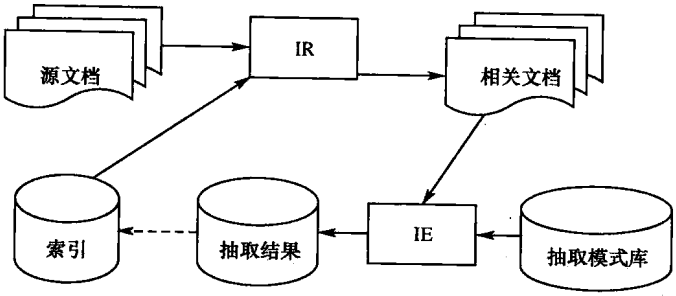


图 1.6 IE 与 IR 互为支持的示意图

首先，它可以应用于传统的信息检索系统之中，在信息检索之后对相关的文本进行指定信息的抽取，使单纯的信息查找过程进一步变成信息匹配（理解）过程，从而把传统的信息检索系统变成智能系统，以用户更满意的方式输出信息。例如，大型联机检索系统、图书情报检索系统、Internet 网页搜寻引擎等，都可以通过引进信息抽取技术来进行改进，或把信息抽取作为辅助功能供用户选用。

其次，信息抽取技术也可以集成到一些数据库应用系统（如产品介绍信息、公司机构名录、分类广告信息、光盘图书报刊阅读等应用系统）中，作为由非结构化的文本自动生成结构化数据（如 XML 数据）的前端工具，使用户能够快速方便地抽取某类指定信息。

3. 从工程角度看，IE 将对以下应用领域产生重要作用

（1）情报收集。情报部门通常需要监控收集敌对国家、网络舆情、恐怖组织活动的各种文献资料。传统上，这种工作需要雇佣大量专门人员阅读、分析和整理。IE 的引入，有利于自动化情报监控，及时发现热点事件和焦点事件。

（2）科技文献监控。IE 系统也可用来辅助进行科技文献的监控任务，获取某学科或技术领域的研究进展情况。例如，对于芯片工艺一些技术参数的抽取，将为有关研究和开发提供有效的支持。

（3）医疗保健服务。医疗保健机构以及健康保险部门可以利用信息抽取系统，获取病人的症状、诊断情况、化验结果以及治疗情况，以便更好地提供医疗服务和保险服务。

（4）商业信息抽取。可以设计专门的 IE 系统，分析新闻中的商业信息，抽取诸如有关公司的合并、合并的参与方以及合并涉及的金额等信息，提供决策支持信息。

1.6 信息抽取的研究现状

1.6.1 国外研究现状

自 20 世纪 90 年代末至今，欧美的主要发达国家组织了许多与知识技术相关的研究项目。来自于自然语言处理（NLP）、人类语言技术（HLT）、计算机语言学（CL）、知识工程（KE）、知识管理（KM）、语义网络（semantic Web）、智能代理（agent based computing）、Web 智能（Web intelligence）以及数字图书馆界的研究组织，围绕着知识表示、知识获取、知识建模、知识抽取、知识检索、知识重用等核心知识技术，开展了一系列以“知识技术”为主题的研究。这些项目一开始就面对实际应用中的信息处理问题，一些研究成果，特别是一些有关知识获取和知识抽取技术的研究成果，早已被一些情报研究机构（如美国国家安全局、美国中央情报局）所采用，在涉及恐怖活动、国际风险投资、竞争情报等领域的情报研究分析和决策咨询中发挥了重要的作用。在这些项目中，卓有成效的有英国的 AKT（Advanced Knowledge Technologies）项目、欧洲联盟（简称欧盟）的 SEKT（Semantically Enabled Knowledge Technologies）和 Knowledge Web，以及美国 DARPA 的项目和 HALO 项目。

1. AKT 项目

AKT 项目的目标是开发和提供一系列技术来解决知识工程和知识管理领域的六个基础瓶颈，其中每个瓶颈都发生在知识演变过程中的一个重要阶段，包括知识获取、知识建模、知识重用、知识检索、知识发布和知识维护。AKT 同时正在为很多领域建立知识门户，包括计算机科学、工程、医学和新闻服务。这些门户包含了用来从我们可以采集到的信息自动链接大量内容的方法，对于协作决策的支持、对于共同存储的帮助，我们可以通过它们来鉴别实用的知识。

AKT 项目目前面对的非常有挑战性的研究问题有：如何实现内容自动或者半自动的被采集和获取；如何克服标引的瓶颈；如何了解一个用户所处的环境从而提供恰当的内容；如何确定内容的质量和出处。

2. SEKT 项目

SEKT 项目的目标是占据欧盟 IT 业领域的核心位置，通过研发核心的语义知识技术来实现欧洲知识社会（European Knowledge Society）。具体进行的研究有：基础研究，自动或半自动本体的产生和增长，本体管理（调解、进化和推论），创新技术的研发，一套知识存取工具，开源的本体中间件平台，通过三个案例研究和基准可用性活动来进行的验证，如何被一种方法论支持的研究。

SEKT 项目的三个核心技术是：基于本体的元数据、人类语言技术和知识发现。这三者必须结合使用。主要的研发挑战有：改善本体和元数据的产生的自动化操作过程；研究和开发本体管理和演变的技术；提供高度可扩展解决方案；研究基于不协调模型的有效的推理机制；开发语义知识存取工具；开发用于配置部署的方法论。

3. DARPA 的 RKF 项目

DARPA 的 RKF（快速知识形成）项目的核心目标是可以使主题相关专家直接便利地进入知识领域并且对知识进行修改，而不需要专门的知识表示、知识获取和知识操作的培训。最终这项技术的目的是保证科研、技术的和军事的专家把大量的知识编码成为可重用的知识库，这些知识库可以用在很多不同的场合下来解决问题。

RKF 项目使用了两个战略性的研究方法：第一个由 Cycorp 领导，使用了一个大型的（百万条公理）按照背景知识组织的知识库，这个知识库是基于 Cycorp 的 CycKB，它包括了一个用于知识录入的自然语言对话系统；第二个由 SRI 领导，使用了一个得克萨斯大学的新方法，通过通用组件的混合和特殊化处理来创建知识，它还使用了一项图形录入技术。更多信息请看 http://reliant.teknowledge.com/RKF。

4. HALO 项目

HALO 项目的目标是成为一个可以包含世界上很多科学知识并且可以回答新鲜问题和提供高级问题解决方案的应用。Vulcan 研究所目前确定的两个主要功能：第一，成为一名可以在科研领域指导学生的老师；第二，成为一名拥有丰富跨学科技能的研究助理，从而帮助科学家进行研究。

HALO 领航员项目是 HALO 项目的第一阶段。HALO 项目的第一阶段证明，知识表示和推理方面的造诣已经能够完成问题回答技术，这项技术目前可以用来回答应用化学领域的新鲜问题并且提供领域相关的解释。

这个项目还证明了当前研究现状的两个缺点：第一，如果每个领域要构建强壮的知识公式，知识的公式化需要高度专业化人员来完成；第二，大部分的系统错误都是由于专业人员缺乏足够的领域知识引起的。

第二阶段通过以下方式来继续推进这两个主要功能：开发一种技术使领域专家能够在逐渐减少对知识工程师依赖的情况下，进行知识的公式化，并且能够找出这些系统的问题。Vulcan 研究所相信，在这个目标上的成功将会把知识公式化的成本减少到可以和编写教科书相差不多的程度，那么此举将会促进科学家、教育家建立一个不断膨胀的机器处理的知识库。第二阶段在五个主要领域达到世

界级技术水平，这五个领域是：知识表示和推理（KRR）、知识获取（KA）、包含自然语言理解的智能接口、可用性和系统整合。更多信息请见 http://www.projecthalo.com。表 1.2 列出了国外几个著名的信息抽取应用系统。

表 1.2 几个著名的信息抽取应用系统

系统名称	所属公司	基本功能和基本情况
InfoXtract	Cymfony	NE、TE、TR 扩展，非受限领域事件抽取，支持开放领域的 QA
SIFT	BBN	NE、TE、TR，完全采用统计方法，训练句子级模型
FASTUS	SRI	核体的瀑布模型，有限状态方法（主要利用模板匹配）
LaSIE-Ⅱ	Shiffield	图形界面，模块化的形式，各模块可以自由组合

1.6.2 国内研究现状

中文信息抽取方面的研究起步较晚，主要的研究集中在对中文命名实体识别方面，在实际实现完整的中文信息抽取系统方面还处在探索阶段。其中，“台湾大学”和新加坡肯特岗数字实验室（Kent Ridge Digital Labs）参加了 MUC-7 中文命名实体识别任务的测评，取得了与英语命名实体识别系统相近的性能。Intel 中国研究中心的 Zhang Yimin 等在 ACL-2000 上演示了它们开发的一个中文命名实体以及这些实体间相互关系的信息抽取系统，该系统基于记忆的学习算法获取规则用以抽取命名实体及它们之间的关系。

在实体关系的抽取方面，姜吉发研究了一种自举的二元关系获取方法，该方法从种子集合出发，获取任意给定的二元关系；哈尔滨工业大学车万翔等参加了 ACE 2004 的实体关系测评，利用 ACE 的训练数据，分别对 SVM 模型、Wino W 算法进行了训练，进行特征选择，并以此进行实体关系的自动抽取，其 F 值均达到了 73%。

北京大学对《人民日报》中的会议消息进行了抽取。山西大学郑家恒等利用最长公共子串，通过聚类的方法对关于农作物的品种描述模式获取进行了研究，取得了良好的结果。

袁毓林在对支持信息抽取的知识资源建设中提出，要有三个层面的语义知识支持信息抽取任务：宏观层的篇章知识，包括段落、小节、句群、句子之间的语义关系；中观层的论元结构知识，包括句子中的谓词和有价名词及其从属成分支配和依存关系；微观层的逻辑结构知识，包括句子中的否定、量化、模态、时体等成分和其所约束的成分之间的语义关系等。

1.7 存在的问题及解决策略

IE研究的最终目标是建立具有较高性能和较好可移植性的IE系统。但是，到目前为止，IE并未和IR一样被广泛应用。原因在于现有IE系统的性能不高，存在如下问题。

1. IE系统中需要的领域相关的模式库和模式匹配功能相分离

前面已指出，按照模式匹配方法实现的一个完整的IE系统由两个大的功能模块组成：模式获取模块和模式匹配模块。前者从一个训练语料中获取模式并将之放到一个模式库中；后者从模式库中取出模式并进行实际的信息抽取。由于采用了这种将模式库从模式匹配功能模块中分离出来的作法，当该系统要从各新领域中进行新任务的IE时，只需将模式库中的模式更新为适合该IE领域任务的模式，而不必修改IE系统的其他功能。这大大地改善了系统的可移植性。

2. 用部分句法分析代替完全句法分析

自由文本中的事件IE模式只能通过语法和语义两个方面来对可能含有事件描述的文本片段进行约束，而语法包括词法和句法。因而无论在学习生成模式的过程中，还是在使用学出的IE模式指导实际IE的过程中，均需要句法分析的支持。但从完成事件IE所需句法信息的层次来看，部分句法分析所能提供的句法信息就足够了。因而目前最新的IE系统都采用部分句法分析来代替完全句法分析。部分句法分析器仅完成对句子中的名词群组、动词群组和介词群组等的识别，因而分析的正确率高、运行速度快。而完全句法分析器的分析正确率较低、运行速度较慢。

3. 采用机器学习方法自动获取IE模式

采用机器学习方法来学习能够指导进行事件IE的领域相关模式规则或统计模型，并不断地改进这些机器学习方法，使得在学习的准备阶段、学习的过程中和学习完成后的模式验证阶段减少用户的工作量并降低对用户的技能要求。最早的模式学习方法需要用户手工标注规模较大的语料；而标注大规模的语料费时、费力，于是为了减轻用户的劳动并降低对他们的技能要求，改进后的模式学习方法只是要求用户将训练文档集合分为相关和不相关的两类，然后就能自动地从中学出相关的IE模式；进一步的研究发现，即使要求用户将训练文档集合分为相关和不相关的两类，也并非易事。因而目前最新的做法是只要求用户提供几个可以轻易想到的有代表性的IE模式，相应的模式学习方法就可以从一个未经分类

的文档集合中学出更多的模式并同时完成对文档的相关性分类。当然，这里我们只是简要地叙述了几种重要的 IE 模式学习方法，更为详细的阐述可以参见 5.3 节。

4. 设计各种跨领域的 IE 模式表达方式

针对从自由文本中进行英文事件的 IE，人们设计了各种各样的模式表达方式。无论这些模式表达方式如何不同，它们都充分利用了语法信息和语义信息的概括约束作用，而且当 IE 系统从一个领域的 IE 转向对另一个领域的 IE 时，这些模式表达方式是固定不变的。

5. 设计图形用户界面

用户通过设计图形用户界面（graphic user interface，GUI）可以方便、快捷地配置 IE 系统所需的领域相关知识，从而便于系统从对一个领域的 IE 转向对另一个领域的 IE。例如，Roman Yangarber 设计的 Proteus 系统的一个领域知识配置界面 PET。

6. 使用领域无关的概念层次知识库的支持

各种事件 IE 模式都利用了语法信息和语义信息的概括约束作用。其中，语义信息的概括约束作用是通过将模式中的某些概念元素用它们的上位概念代替来完成的，而这就需要一个概念层次知识库的支持。这个概念层次知识库由领域相关的概念知识和领域无关的概念知识两部分组成，而领域无关部分的概念知识可以直接采用现成的领域无关的概念层次知识库如 WordNet 等，需要用户手工生成的只是领域相关的概念层次知识库部分，这自然大大减轻了用户在 IE 模式获取过程中的工作量。Joyee YueChai 等设计的 xE 系统 TxMEs 就采用了 WordNet 的支持，对 IE 模式进行语义泛化。

1.8 信息抽取的挑战和趋势

1. 移植 IE 系统面临的问题

信息抽取面临的主要挑战是知识获取的瓶颈问题，即信息抽取的适应性问题。在特定的领域构建信息抽取系统，技术上已基本成熟，但知识的自动获取实际上仍没有达到完全自动，大部分系统只是把原先由领域专家完成的任务转化为用户的任务。在构建通用的知识学习器方面，部分文献进行了有益的探讨，但效果不是很理想。目前，移植 IE 系统面临以下四个方面的问题。

（1）适应新的领域信息。构建系统资源（如词库、知识库等），并设计新的

模板使系统可以处理一些特定领域的概念。

（2）适应不同子语言特征。修改语法和词库，使系统能处理应用或领域内典型的特定语言结构。

（3）适应不同的文本流派。特定流派的文本（如医学结论、科学论文、政策报告等）具有特定的词汇、语法和篇章结构。

（4）适应不同类别的文本。基于 Web 的文档可能与新闻报纸之类的文本有着强烈的差别，必须能适应不同的情况。

2. 信息抽取呈现的趋势

（1）信息抽取的范围不断扩大。从信息抽取的信息源看，早期的信息抽取主要集中于自由文本，现在的信息抽取则扩展到话语信息抽取和 Web 页面信息抽取。话语文本分析不同于一般的文本，当话语转换为文本时，会出现信息的增加和丢失（识别错误引起）现象，信息抽取技术也必须能适应这种现象。Web 页面作为海量的信息存储所，近年来尤其受到信息抽取和文本挖掘技术的关注。从信息抽取的领域看，已从军事、政治、医学等领域，扩散到商业、科技等领域，且仍有进一步扩大的趋势。

（2）信息抽取技术的多样化。信息抽取一般与领域性知识有较紧密的关系，因此，最初的信息抽取与子语言的处理技术也极为相似，正规文法、上下文无关文法和自动机技术等应用得较为广泛。随着语料库的成功构建，特别是 Web 页面的迅猛增长，基于统计的技术和机器学习方法在信息抽取方面发挥着越来越重要的作用。可以说，信息抽取技术已摆脱了狭义的自然语言理解技术的束缚，向着多样化的方向发展。

（3）知识获取的进一步自动化。信息抽取面临的主要挑战是系统在领域间的可移植性问题，这一问题关系着信息抽取技术适用范围的大小。知识的自动获取就是针对这个问题而提出的，并经历了手工编码、半自动获取和自动获取三个发展阶段，知识的自动获取已成为信息抽取技术的核心。目前，知识的获取主要面临三个方面的问题：①没有提出标准的知识框架。领域之间所需知识差别很大，通用的知识框架能帮助快速获取这些知识。在这方面，概念节点是个很好的范例，但没有作为标准提出。②知识的自动获取范围较窄。目前仅限于规则模式的自动获取，而对于如 CRYSTAL 等至关重要的概念层次等仍由手工编码完成。③自动化的程度仍偏低，要求一定的手工参与。覆盖性算法等虽然要求用例较少，但对所用实例一般要求较高，必须细心选择。因此，知识获取自动化仍是研究的重点。

3. 当前 IE 研究的直接目标

IE 研究的最终目标是建立具有较高性能和较好可移植性的 IE 系统。但是，到目前为止，IE 并未和 IR 一样被广泛应用。原因在于现有 IE 系统的性能不高、可移植性不好。在 MUC 举行的多次针对不同领域的具有不同复杂度的事件 IE 任务测试中，很少有 F 超过 0.60 的 IE 系统。由于不同的 IE 领域任务需要与不同领域任务相关知识库的支持，因而 IE 系统从本质上是领域任务相关的。所以，如果对每个 IE 领域任务都开发相应的 IE 系统来应对的话，代价就太昂贵了。

要提高 IE 技术应用的广度，即将应用在各种不同的信息处理技术中，应该进一步提高 IE 系统的性能和可移植性。

实际的事件 IE 过程由一系列的 NLP 步骤组成，并且需要领域任务相关知识库的支持，所以 IE 系统的最终性能便依赖于各个 NLP 步骤的性能和所需领域任务相关知识库的全面性和正确性。让 IE 研究者自己去改善各个 NLP 步骤的性能是困难的，这需要其他计算语言学家的支持才行。而且，在许多具体的 IE 应用背景下，一个性能不高的 IE 系统就足够了。所以，近些年来大部分的 IE 研究者都把改善 IE 系统的可移植性作为其直接研究的目标。

第 2 章　信息抽取评估

2.1　信息抽取评估一般原则

一个适应性的信息抽取系统应该具备以下要求：准确性、弹性、自我修复性、通用性、可扩展性、开放性。

从以上要求来看，目前大多数的自动信息抽取系统只能处理有限的文档格式，不能很好地适应文档结构的变化，适应性有限。

对信息抽取系统的评估可以从三维的角度去进行：信息抽取任务的复杂度、系统所使用的技术、系统的可移植性评估（手动建构的、半自动的或完全自动的信息抽取系统）。

1. 抽取任务复杂度的评估

第一维度的评价实际上可以回答为什么信息抽取系统在处理具有特殊结构的网页时会失败。信息抽取任务的难易主要取决于以下三个因素。

（1）文本的类型。信息抽取的文本既可能来自于学术期刊，也可能来自 WWW 上的 HTML 文档，或电子邮件信息。不同来源的文本有不同的格式规范，有时信息抽取可利用这些格式上的隐性信息。

（2）涉及的领域。信息抽取应用得较为广泛，既可用于金融领域，也可用于旅游业或技术支持性的领域等。

（3）抽取的场景。既有公司、企业之间合并信息的抽取，也有关于恐怖主义活动等的信息抽取。

因此，抽取任务复杂度的评估必须兼顾文本类型、涉及领域、抽取场景的多态性。在所有的信息抽取中，实体抽取是最底层的任务，也是必不可少的。因此，抽取任务复杂度可从文本中涉及的实体关系方面进行分析。首先将抽取的任务用实体网络进行描述，在实体网络中，以节点表示实体，以弧表示实体之间存在的关系。一个实体网络中含有的节点和弧越多，表明相应的抽取任务越复杂。因此，弧的数目在一定意义上表明了抽取任务的复杂度。

2. 信息抽取系统性能的评估

第二维度是通过比较各种系统所利用的技术来评估。信息抽取系统产生固定

有关此电子书的说明

本人可以帮助你找到你要的PDF电子书，计算机类，文学，艺术，设计，医学，理学，经济，金融等等。质量都很清晰，为方便读者阅读观看，每本100%都带可跳转的书签索引和目录，只要您提供给我书的相关信息，一般我都能找到，如果您有需求，请联系我 QQ1779903665。

PDF代找说明：

本人已经帮助了上万人找到了他们需要的PDF，其实网上有很多PDF,大家如果在网上不到的话，可以联系我QQ，大部分我都可以找到，而且每本100%带书签索引目录。因PDF电子书都有版权，请不要随意传播，如果您有经济购买能力，请尽量购买正版。

提供各种书籍的pd电子版代找服务，如果你找不到自己想要的书的pdf电子版，我们可以帮您找到，如有需要，请联系 QQ 1779903665.

备用:QQ 461573687

声明：本人只提供代找服务，每本100%索引书签和目录，因寻找和后期制作pdf电子书有一定难度，仅收取代找费用。如因PDF产生的版权纠纷，与本人无关，我们仅仅只是帮助你寻找到你要的pdf而已。