中国科技资源导刊 ISSN 1674-1544
2008年3月 第40卷第2期 52-56
CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
ISSN1674-1544 Vol.40 No.2 52-56,Mar. 2008

信息抽取技术及其在数字图书馆中的应用前景

李 昕
(淮安市行政学院,江苏淮安 223001)

摘 要: 信息抽取的目标是自动从文本信息中抽取出预先想要得到的信息(知识),它提供了一条从浩瀚信息堆积中抽取出与用户相关信息的一条思路。本文分析了信息抽取的主要概念、信息抽取的现状及类型,提出了在数字图书馆的建设中,信息抽取技术在数字内容的自动标引、数据获取、数据挖掘、情报研究分析、参考咨询等方面发挥重要的作用。

关键词: 信息抽取;数字图书馆;应用前景
中图分类号: G250.76 **文献标识码**: A DOI: 10.3772/j.issn.1674-1544.2008.02.011

A Prospect Analysis of the Application of Information Extraction in Digital Libraries

Li Xin
(Huai'an Administrative College, Huai'an, 223001)

Abstract: The aim of information extraction is to automatically get expected knowledge from texts, which provides a new method for users to get useful information easily. This article first analyses the main concepts of information extraction, the current situation and types of information extraction, then puts forward that in the construction of digital libraries, information extraction technology can play siginificant role in automatic indexing of digital content, data acquirement, data mining, information research, and information consultation.

Keywords: information extraction, digital libraries, application prospect

任何一项技术的研究都不能孤立存在,需要相关技术的补充与推动。自然语言处理技术、人工智能技术、语言工程技术的发展推动了信息抽取技术走向成熟。当前国外对信息抽取技术研究较多,例如 AVENTINUS、E-CRAN、GATE、LaSIE 等研究项目,长期对信息抽取及自然语言处理技术进行跟踪研究,同时也形成了如 AN-NIE、MELITA、AMILCARE 等较成熟的系统。

随着数字图书馆采集技术、存储技术和检索技术的进步,数字图书馆的技术瓶颈日益集中在数字信息的深层次分析处理上。面对数量日益庞大的数字信息堆积,如何高效地对这些数字信息进行加工处理,有效地实现数字资源

第一作者简介:李昕(1972-),女,江苏宿迁人,副研究馆员,主要研究方向是图书馆理论及计算机网络。

收稿日期:2008年1月8日。

的开发利用已经成为当前数字图书馆研究者必须面对的一个课题。

1 信息抽取技术的涵义和类型

信息抽取是从一段文本中抽取指定的一类信息(例如事件、事实),并将其形成结构化的数据填入一个数据库中供用户查询使用的过程。其主要目标是让计算机不但找到相关的文档,而且还能找到相关的内容。例如,从关于计算机的文本中抽取设备名字、用途等特定信息。一个典型的信息抽取(Information Extraction, IE)任务是从在线文本中抽取相关的信息,填写到预先定义好的模版中的属性槽里。这种任务的主要优点在于有效地过滤掉当前文本与特定领域无关信息,而深入的自然语言处理技术必须对整个文本进行完全分析。

特定领域的信息抽取系统任务与通用的自然语言理解任务不同[1]。对于通用的自然语言理解来说,系统必须对输入的句子进行深入分析,产生包含输入句子所有意义(包括隐含意义)的表达。一般来说,理解分为两步:第一步通过句法分析将输入的句子映射到一个句法结构中,如句法树;第二步,通过句法到语义的转换分析实现将句法结构映射到意义表达。而对于特定领域的IE来说,是从一段文本中抽取指定的一类信息并将其形成结构化的数据填入一个数据库中供用户查询使用。由于需要抽取的信息类型也是预先定义好的,在相关的句子中,只有一些携带相关信息的短语单元才能被解释,输入的文本只能映射到一些有限数目的事件分类,如关于爆炸事件、凶杀事件等,所以完全句法分析和深入的语义解释没有必要。不必对文章的全部内容进行全面的分析,而只是对有用的文章片段进行有限深度的分析,其效率与灵活性都比较高。

信息抽取以模版框架[2]为中枢,分为选择与生成两个阶段。模版框架是一张申请单,它以空槽的形式抽取应从原文中获取的各项内容。例如,针对计算机病毒类的文章可以提出如下的分类:病毒名称,病毒传染对象,病毒类属,病毒攻击对象等。在选择阶段,利用特征词从文本中抽取相关的短语或句子填充模版框架。例如,在文本中发现"……感染可执行文件……"字样,则可以将特征词"感染"后面的短语"可执行文件"作为病毒的感染对象填入模版框架。模版框架是带有空白部分的现成的套话,其空白部分与模版框架中的空槽相对应。例如,"该病毒的感染对象是(病毒传染对象)"是模版中的一个句子,因为在模版框架中登记的病毒感染对象为"可执行文件",因此在信息抽取中将输出"该病毒的传染对象是可执行文件"句子。

信息抽取技术结合了自然语言处理、语料资源以及语义等技术,目的是从无结构的自由文档或其他资源中抽出结构化的无二义信息,它是元数据抽取、信息分析、信息索引及检索的基础。信息抽取系统(Information Extraction System, IES)的主要功能是从文本中抽取出特定的真实信息。被抽取出来的信息以结构化的形式描述,直接存入数据库中,供用户查询检索以及进一步分析。为了处理海量文本,信息抽取系统以信息检索系统的输入作为输出,信息抽取技术提高了信息检索系统的性能,这样的结合能更好地服务于用户的信息处理需求。广义的信息抽取系统的处理对象可以是语音、图像、视频等其他媒体类型的数据。狭义的信息抽取研究,即针对语言文本的信息抽取,在本文中主要探讨狭义的信息抽取技术。典型的信息结构系统结构被抽象为"级联的转换器或模块集合,利用手工编制或自动获得的规则在每一步过滤掉不相关的信息,增加新的结构信息"。信息抽取基本系统结构见图1。

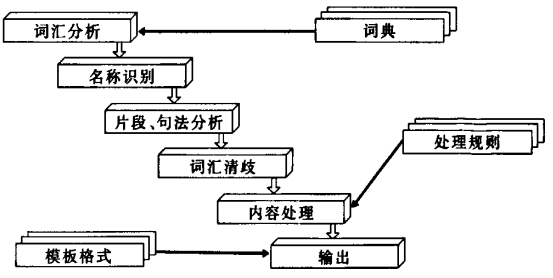

图1 信息抽取基本系统结构

根据信息抽取内容以及所抽取出的信息的集聚水平的不一样，将信息抽取分为以下几种主要类型。

(1) NE: 即命名实体识别 (Named Entity Recognition)，它是信息抽取中最为基本的任务，它仅需要系统能够从众多信息中标识并分离出相关的命名实体。此类信息抽取需要系统能够识别出实体名，并将相应的实体名进行归类。信息处理会议 (Message Understanding Conference, MUC) 测评需要信息抽取系统能够从自由文本中识别并抽取出人名、组织名、日期、时间、地点以及某种类型的数字表达式 (如货币数量，百分数)，并在文本中对这些信息进行标注。命名实体识别具有非常直接的实用价值，在对文本中的名称、地点、日期等进行标注之后，即提供了对这些信息进行检索的可能。对于许多语言处理系统，命名实体识别是其中一个很重要的组件，是目前最具实用价值的一项技术。

(2) MET：即多语种实体识别任务 (Multi lingual Entity Task) 信息抽取。MET 除了能够对英文命名实体进行识别之外，还需要能够对多语种的命名实体进行识别，例如可以对中文、日文或西班牙文进行命名实体识别。

(3) TE：即模板元素 (Template Element) 信息抽取，它将特定的描述信息与实体联系起来。它需要从文本的任何地方将与组织、人物或其他实体相关的基本信息抽取出来，并将这些信息作为实体的属性进行聚集，形成实体对象。在 MUC 评测中，TE 系统需要能够从文本中抽取特定类型的实体信息，并将这些信息填写到预先定义的小型的属性模板之中。例如，对人物实体的模板元素抽取，需要信息抽取系统能够抽取出预先定义的人物的名称、职务、国籍等属性。

(4) CO：参照 (Coreference) 涉及在进行 NE 或 TE 任务时，从文本中标识出对同一实体的不同表达方式。例如，连接某同一实体的不同称谓，将某一名词和其相应的代名词进行连接。在 MUC 中，CO 之所以得到重视，是因为它能够为创建 TE 和 ST(见下文)打下基础。CO 可以将散布在文本中不同地方的同一实体的描述信息连接起来，同时通过分析实体在文本中不同地方出现的情况，以及此实体在不同场合与其他实体之间的关系，有助于情节信息的抽取。

(5) TR：即模板关系 (Template Relation) 信息抽取。TR 需要在 TE 的基础之上标识出模板元素之间的关系。例如职工和组织之间的关系、产品和生产企业之间的关系以及公司和地区之间的关系等。TR 是信息处理会议第七阶段 (Message Understanding Conference 7, MUC7) 定义的一项新任务，它的抽取包括相关元素模板，以及元素模板之间的相互关系。

(6) ST：即情节模板 (Scenario Template) 信息抽取。ST 抽取某一事件中的事件信息并将事件信息与某个组织、人物或其他实体相关联。ST 需要标识出特定事件及事件的相关属性，包括将事件中的各个实体填充到事件的相应角色中，通过各个对象之间的关系，能够还原出整个事件的“原型”。

2 信息抽取技术在数字图书馆中的应用前景

2.1 构建模板挖掘平台

2.1.1 自动建立数字文献的引文数据库

通过快速而简单地审查一些数字图书馆的电子期刊发现，模板挖掘可用于自动地从在线文章中建立引文数据库，这种引文数据库可在后来被用于各种引文分析和其他目的。这些数据库包含 ISI 数据库熟悉的一些信息，例如引用作者、引用作者的地址、引用论文的标题、关键词以及作者、标题、引用论文的书目信息等。

2.1.2 自动识别用于研究资金、赞助机构

通过浏览文章中的“致谢”一节，可以发现它们包含研究的资金、赞助机构的名字、地址、授权号等。因此，对电子期刊中各种文章的研究能帮助产生一种模式，通过这种模式形成适当的模板，而这种模板可以在更多的应用中抽取相关的信息。

2.1.3 利用元数据和模板挖掘抽取信息

元数据是从数字文献中发现资源的重要工具。它们不仅帮助用户定位需要的信息资源，也能帮助检查、选择(或拒绝)检索条目。元数据标记，例如都柏林核心集(Dublin Core, DC)元数据格式的主题和描述，可以被扩展并形成各种特定的结构化的模板。例如，有专家提出用预组配标题索引法 (Pre - coordinate Indexing) 原理建立主题模板来进行信息抽取。用模板挖掘方法结合元数据格式能更好地进行信息检索，并能自动产生文献代理来帮助最终用户挑选输出信息。这对网络环境下图书馆的信息检索非常必要。

2.2 构建情报自动搜集平台[3]

情报研究分析需要有大量的相关信息作基础，而数字图书馆中的信息搜集面向大量的Web网页。

一般的Web网页需要抽取4种类型知识：基本知识、块模式知识、确定模式知识、非确定模式知识。针对这4种抽取知识类型，利用网页智能搜索技术和网页内容自动抽取技术，可从诸多门户网站中，自动搜索出感兴趣的网页，并从中自动识别和抽取所需要的信息内容。情报自动搜集平台可以由三大模块组成(图2)。

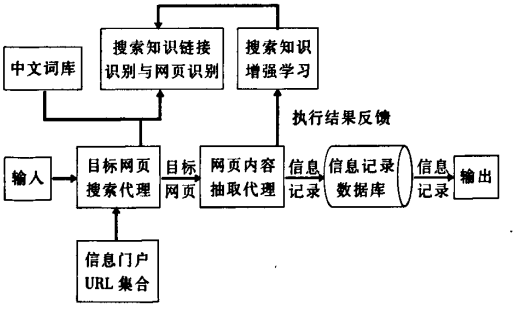

图2 情报自动搜集平台架构

2.2.1 基于最佳搜索路径的目标网页搜索代理

该模块利用中文词库，目标网页相关搜索知识，对信息门户URL集合中各个门户网站进行智能搜索，快速搜索出所需要的目标网页。其中目标网页搜索知识主要包括两部分：最佳搜索路径上的网页链接识别知识和目标网页识别知识。

2.2.2 网页内容抽取代理

该模块接受目标网页搜索代理搜索获得的目标网页，对网页内容的数据记录表示格式，进行自动学习归纳，并最终自动获取记录信息描述模式，然后根据这些记录信息描述模式，将目标网页中所需要的记录信息抽取出来；将这些抽取出来的信息记录存放在相应的信息记录数据库中，以提供给用户查询。

2.3.3 搜索知识增强学习模块

该模块接受网页内容抽取结果成功与否的结果，将所抽取的网页作为目标网页搜索相应的正反实例，并利用增强学习来完成目标网页的搜索知识[4]。

2.3 数字内容的自动标引和元数据获取

数字图书馆面对海量的信息资源，需要提供有效的信息检索和内容揭示方式，内容标引和元数据加工是数字图书馆区别于其他低品质信息检索系统的一个重要方面。但是，随着信息资源的迅猛扩增，手工的内容标引和元数据加工已远不能适应这一需要。探索有效的内容标引和元数据抽取成为一个十分迫切的问题。我们知道，对信息抽取的一种理解就是从自由文本中抽取格式化的信息，它可以作为一个从无结构的自由文本或其他信息资源中抽取出结构化的、无二义性信息资源的过程。实际上，国外的一些研究人员已经注意到了信息抽取在数字图书馆元数据建设方面的作用，出现了将信息抽取技术应用于数字图书馆的内容标引和元数据抽取的相关研究。相信随着研究的深入，会出现更具有实用价值的成果。

2.4 参考咨询中的问题解答

其实，问题解答(Question Answering, QA)也是自然语言处理 (Natural Language Processing,

NLP)研究中的一项重要内容。问题解答系统能够让用户以自然语言的方式提出问题，系统通过对大量相关数据的查找、分析和推理，从知识库中整理出针对这一问题的答案。数字图书馆中的参考咨询正在促进数字图书馆服务方式从检索方式到问题解答方式的转变，目前的参考咨询系统主要凭借馆员个人的学识对读者的问题进行解答。然而NLP技术的进步，已经开始显示出自动从知识库中获得答案的可能。已经有研究表明，信息抽取技术能够支持问题解答系统。

2.5 情报研究分析

数字图书馆的情报研究分析亦需要从大量的相关信息中研究分析出事件发展的各种态势。大量的数据和相关信息是进行研究分析的基础，但这些信息和数据从何处而来？信息抽取提供了一条进行大规模数据及信息采集的思路。通过信息抽取，能够从自由文本中抽取出数值数据和结构化的信息，建立起可供研究分析的联机分析系统，进而实现大规模的数据挖掘和信息分析。

2.6 大型知识库、数值库建设[6]

数字图书馆长远目标是从信息检索服务转向知识提供服务。知识提供的前提是知识的获取。如何有效地获取知识呢？目前，许多研究人员将人类语言工程(HLT)和知识获取结合起来，提出了通过Ontology驱动的信息抽取来实现知识的获取。还有的提出了从非结构化文本中建设知识库的思路。

参考文献

[1]朱谨波，等．中文信息自动抽取[J]．东北大学学报，1998(1)：52.

[2]王宁．模版处理在信息抽取过程中的应用[J]．江西图书馆学刊，2001(2)：50－52.

[3]王凯，等．信息抽取系统在高校数字图书馆的应用[J]．现代情报，2006(4)：87.

[4]朱明，等．基于主题的Web信息个性化服务[J]．计算机应用．2002(12)：23.

[5]朱明，等．基于Web的企业竞争对手情报自动搜集平台[J]．微计算机应用．2004(1)：46.

[6]刘爽．信息抽取技术及其在数字图书馆中的应用前景分析．现代情报，2006(11)：76.