Web信息抽取*)

李 晶 陈恩红
(中国科学技术大学计算机系 合肥230027)

Web Information Extraction

LI Jing CHEN En-Hong
(Department of Computer, University of Science and Technology of China, Hefei 230027)

Abstract With the tremendous amount of information available on the Web, the ability to quickly obtain information has become a crucial problem. It is not enough for us to acquire information only with Web information retrieval technology. Therefore more and more people pay attention to Web information extraction technology. This paper first introduces some concepts of information extraction technology, then introduces and analyzes several typical Web information extraction methods based on the differences in extraction patterns.

Keywords Information extraction, Information retrieval, Semi-structured text, Extraction pattern, XML

1 引言

自九十年代初互联网开始迅速发展至今,互联网已成为经济、社会、文化、教育以及娱乐等各个方面的重要组成部分,并正在成为我们工作和生活中不可或缺的一员。随着互联网的迅速发展,互联网上流通的信息也在爆炸性地增长。为帮助互联网用户有效地发布与接受信息,众多的互联网搜索引擎如Yahoo、Excite和AltaVista等不断涌现,向广大互联网用户提供基本的信息搜索服务。但进入九十年代后期,随着互联网开始步入正常发展阶段,仅仅靠搜索引擎已越来越难满足人们对互联网信息服务的需求,因为它们所能覆盖的网页占整个互联网网页总量的比例越来越小,更主要的问题是,随着互联网搜索引擎所覆盖网页的不断增加,互联网用户将会发现越来越难有效地利用这些搜索引擎,来帮助自己发现所需要的互联网信息资源。面对浩如烟海的互联网信息资源,仅仅依靠浏览器以及基于关键字检索查询的搜索引擎,已远不能满足互联网用户的信息需求。如何帮助人们准确有效地找出自己所需要的信息资料,已越来越迫切地摆在了我们的面前。本文介绍的Web信息抽取,就是解决如何准确有效方便地从Web网页中抽取出所需要信息内容的一项技术。

2 信息抽取概述

信息抽取(Information Extraction)就是从一段文本中抽取指定的一类信息(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。例如从一篇关于自然灾害的新闻报道中抽取出灾害的类型、时间、地点、人员伤亡、经济损失等情况。信息抽取系统的实现一般有知识工程方法和自动训练方法。所涉及的技术包括:自然语言处理技术、人工智能技术、语言工程技术等。

信息抽取迄今已在很多领域取得了成功应用,例如:超市分析交易数据,安排货架上货物摆布,以提高销售;调查局分析行为模式,判断哪些人对受保护的信息具有潜在威胁;保险公司分析以前的客户记录,决定哪些客户是潜在花费昂贵的;医师分析病人历史和当前用药情况,不仅诊断用药而且预测潜在的问题;税务局分析不同团体的交所得税的记录,发现异常模型和趋势;分析有线新闻和广播电视的文本来找到和总结恐怖分子活动记述;监控微电子芯片制作的技术报道,从而捕捉芯片销售、制作技术进步和芯片处理技术发展或使用情况等方面的信息。

信息抽取系统利用一种由事件名称(Event)、日期(date)、时间(time)、地点(location)等槽(slot)组成的信息模式,对报道中相应的内容进行匹配,并正确填满各槽的内容。一般而言,一个典型的信息抽取系统的工作过程主要包括了如下几个步骤[1]:

1. 用一组信息模式描述感兴趣的信息。系统可以针对某一领域的信息特征预定义好一系列的信息模式,存放在模式库中供用户选用。

2. 对文本进行"适度的"词法、句法及语义分析,通常包括识别特定的名词短语(人名、机构名、产品名、事件、地点等)和动词短语(事件描述、事实陈述)。这需要使用合适的词典、构词规则库等知识库的支持。

3. 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)。

4. 进行上下文关联、指代、引用等分析和推理,确定信息的最终形式。

5. 输出结果(例如生成一个关系数据库或给出自然语句陈述等)。

出于效率的考虑,有的信息提取系统还包括一个预处理过程,目的在于过滤掉与提取目标不相关的文本。

信息抽取评估标准源于信息检索的两个评估标准:查准率(Precision)和查全率(Recall)。然而信息抽取的这两个标准尽管名称和信息检索的一样,但意义已有所不同[1,2]:

(1)查准率等于系统产生正确答案的数目除以系统产生的所有答案的数目;

*)本文研究得到安徽省自然科学基金(01042302)资助。李 晶 硕士生,主要研究领域为信息抽取与数据挖掘。陈恩红 博士,副教授,主要研究领域为机器学习、数据挖掘与智能信息处理等。

(2)查全率等于系统产生正确答案的数目除以文本中所有可能的答案数目(包括系统得到的和系统不应该忽略的)。

查准率主要测试系统的准确程度,即描述系统检索或抽取的信息中,有用的是多少。查全率主要测试系统的理解程度,即表示应该得到的信息中,已查出了多少。通常,我们借助于这两个指标来衡量信息抽取系统的优劣。查准率与查全率的值都在[0,1],它们最理想值是1.0,但几乎无法达到,据统计,目前较复杂的信息抽取任务查准率最高的可达70%,查全率可达50%,而那些比较简单的信息抽取任务查准率和查全率最高均可达到90%。

查准率和查全率并不是互不相关的,而是相互影响的,也就是说期望得到较高的查准率,那么得到的查全率就会低一点,反之亦然。当我们在比较不同信息抽取系统的性能时,应该同时考虑查准率和查全率。这里有一个公式:

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

注:P 指 precision 值,R 指 recall 值,参数 β 决定了对查准率和查全率的支持度。通常取 β 值为1,表示对 precision 和 recall 的支持度相等。采用该公式,对于查准率和查全率值不同的系统就很容易进行比较了。

3 Web 信息抽取方法

Web 是一个巨大的信息源,其上的信息可以分为3类:自由文本、结构化文本、半结构化文本,但以半结构化文本为主。针对这三种文本,信息抽取也分为三种类型[3,4],它们分别是:(1)从自由格式的文本中抽取出所需要的信息内容;(2)从半结构化的文本中,抽取出所需要的信息内容;(3)从结构化的文本中抽取出所需要的信息内容。其中第一种信息抽取工作最为困难,而第3种信息抽取任务最为简单,结构化文本是指数据库中文本信息或遵循预先定义的而且严格的格式的文本,这样的信息易于使用格式描述进行抽取。对这种事先格式已知的文本进行信息抽取通常用比较简单的技术就可以了。而第二种文本即半结构化文本是介于非结构化文本和完全结构化文本之间的一种类型,以前的信息抽取系统无法有效地处理此类文本,因为此类文本在风格上常常是不合乎文法的,不遵循任何严格格式。对于半结构化文本传统的采用自然语言处理技术的信息抽取系统已经不适用了,其抽取模式经常是基于标记和分界符,如 HTML 标记等,句法和语义信息只是在一定范围内被使用。

对于一个信息抽取系统而言,其关键的一个元素就是抽取模式(抽取规则)的描述,下面就以抽取模式为中心介绍几种具有代表性的 Web 信息抽取系统。

3.1 自由格式文本的信息抽取

这类文本的抽取模式常常都是基于语法和语义的约束描述,这些约束将帮助确认文档中(待抽取)的相关信息。因此为了利用这些抽取模式,文档都必须事先经过语法分析器和语义标志器的处理。

AutoSlog 构造了一个由概念和概念结点组成的抽取模式字典[5]。每个 AutoSlog 概念都包含一个概念链接(Anchor)和一个语言模式,以及一个适用条件集合。这里的概念链接实际就是一个触发单词,而适用条件则表示对语言模式的约束。例如:为了从下面句子中抽取出恐怖分子的攻击目标:"The Parliament was bombed by the guerrillas",可以使用一个包含触发单词"Bombed"的概念,它的语言模式是"(subject) passive-verb"。应用这个模式抽取信息过程就是:首先由于句子包含"bombed"单词,相应概念被激活;接着语言模式用于对句子内容的匹配,subject 的内容被抽取出来作为恐怖分子的攻击目标。AutoSlog 共利用了13个事先定义好的语言模式,所抽取的信息可以是以下几种情况:subject、direct subject、noun phrase。通常触发单词是动词,而被抽取的信息是 noun phrase 时,触发单词也可以是名词。

LIEP 是一个自学习系统,可以实现多槽抽取[6]。也就是说,并不是一个抽取模式的定义只能抽取你所感兴趣的句子中的某一项内容,而是可以抽取所有多个项。抽取模式是由两部分组成:语法约束(如 TRGT 必须是句子的主语,而且该句子包含一个由介词短语修饰的动词)和语义约束(如 TRGT 是一个物理实体,动词使用被动语态,介词短语用"by"开始)。

PALKA 系统的抽取模式是用语义框架模式结构,简称 FP 结构[7]。一个 FP 结构是由一个语义框架和一个短语模式组成。语义框架中的每个槽定义了待抽取的项以及对它的语义约束(例如爆炸事件的目标 target 项必须是一个物理实体)。而短语模式定义了词汇入口的顺序和从事先定义的概念体系中挑选出来的语义种类。FP 结构通过语义框架中的槽和短语模式中的元素对应而将语义框架和短语模式结合起来。将 FP 结构应用到句子中是一个很简单的过程:如果短语模式和句子匹配,那么 FP 结构被激活,相应的语义框架也被准确地用于抽取数据。

与 AutoSlog 不同,FP 结构既可以通过正确的匹配被激活,也可以通过事先定义的概念体系中 is_a()关系来激活,然而有趣的是,PALKA 系统表达力却不如 AutoSlog,因为 FP 结构只对动词有准确的单词约束(exact word constraints),而 AutoSlog 还可以对名词有准确的单词约束。CRYSTAL 构造了由多个槽(slot)组成的概念结点,该概念结点能够对任何构成的短句施加语义和单词约束。CRYSTAL 概念结点的表示方法具有更强的抽取模式知识表达能力。

上述的各个系统都用于对合乎文法的自由文本进行数据抽取。尽管它们每一个都用到语法和语义约束来识别感兴趣的内容项,它们仍然有几个比较重要的不同点。

首先,抽取的粒度不同:LIEP 能够识别正确的短语,而 AutoSlog、PALKA 和 CRYSTAL 只能决定包含目标短语的语法字段(syntactic field)。其次,除了 CRYSTAL,所有其它的系统允许只对被抽取槽进行语义类约束。(其它句子元素允许准确的单词和动词词根约束)。第三,PALKA、CRYSTAL 定义的抽取模式可以是单槽的,也可以是多槽的,而 AutoSlog 的抽取模式只能实现单槽抽取,LIEP 系统不能产生单槽的抽取模式。还有一点,如果被抽取的某项既可以出现在主语位置,也可以出现在介词短语中,那么 AutoSlog、PALKA、CRYSTAL 必须创建两个不同的抽取模式,而 LIEP 则不受这样的限制,只要创建一个抽取模式就可以覆盖两种情况。

3.2 结构、半结构文本的信息抽取

随着互联网的迅速发展,人们开始接触到大量的符合语法和不符合语法的文本信息资源。例如:我们需要从网上招工布告栏中抽取出有关的招工信息,或者从公寓招租布告栏中抽取出有关的租屋信息等等,显然这些信息抽取需求都具有很强的应用背景。自由文本的抽取技术已不适用于在线文本的抽取,为了完成这类信息抽取任务,有关的抽取模式知识就需要既包含语法和语义两方面的约束,也包含用于确认需抽取文本内容的边界。

RAPIER 能够学习单槽抽取模式，只是这个模式知识仅使用了有限的句法知识[8]。模式知识主要包括三部分内容：前(pre)和后(post)-filler patterns 分别定义左右边界，而“Filler pattern”分别指示待抽取信息前面的单词，以及后面的单词。“Filler pattern”指示待抽取的信息最多的单词数且单词的词性。

WHISK 是一个学习系统，它能够自动产生从各种格式文本中抽取有关信息所需的相应模式知识[9]。WHISK 的信息抽取模式知识主要包括两部分内容：描述有关信息的上下文内容(content)，以及描述所需抽取信息的准确边界(delimiter)。根据文本的结构，WHISK 利用上述两部分描述构造相应的信息抽取模式知识。而 SRV 的抽取模式是基于属性值测试和文档的相关结构[10]。STALKER 系统可以从半结构化的网页内容中抽取具有层次结构的信息内容。例如网页内容与饭店有关，由于在一个城市中可能有几个地址，而每个地址可能有几个电话号码。为有效解决这一个多层次嵌套信息的抽取问题，STALKER 采用 Embedded Catalog Tree (ECT)来帮助描述文件的组织结构，以及抽取任务的输出模式。同时也帮助指导信息的抽取过程[11]。给定一个 ECT，STALKER 为每个 ECT 结点产生一个抽取规则，对每个 LIST 结点再产生一个附加循环规则。整个抽取过程也是按照这个层次结构展开。

这四种类型的抽取系统有以下几个方面不同：首先，RAPIER 和 SRV 仅能产生单槽的抽取模式，对于大规模的域具有很大的局限性，例如包括几个姓名和地址的文档中，单槽抽取模式很难识别某个指定人物的地址。另一方面，尽管 WHISK 能产生多槽模式，但是它不能自动将文档分段而将抽取模式只用于文档中的片断；其次，RAPIER 和 SRV 设置的约束集合比 WHISK 更丰富：RAPIER 使用一个 part-of-speech tagger，SRV 利用拼写特征、标记长度和链接语法。而且，这两个系统将约束基于 WORDNET 语义类。第三，如果被抽取的短语长度随文档变化而自动变化，RAPIER 和 SRV 可以抽取太多或太少单词，因为这两个系统可能强制短语长度。第四，STALKER 与前3种抽取系统相比最大的特点就是实现了从半结构化的网页内容中抽取具有层次结构的信息内容。

3.3 基于 KPS 的 HTML 的信息抽取

目前 Web 上文档主要是 HTML 格式的，为此我们在此特别介绍基于 KPS(Keywords、Patterns 和 Samples 三个单词的首字母缩写)的 HTML 的信息抽取[12]。这三种方法可以单独使用，也可以结合起来使用，以进一步提高数据提取精度。

3.3.1 *基于关键字的数据抽取* 其方法的主要思想是：首先分析人们发布信息的日常习惯，建立一套启发式规则，然后根据给定的关键字，在 HTML 文档中查找此关键字，找到后，再应用这些启发式规则，抽取出所需的目标信息。基于关键字的数据抽取方法主要用于抽取跟某个关键字相关的简单数据值，如某人的 email 地址、电话号码等。

以下是一些常用的启发式规则：

·若关键字出现在一个链接的标签里，则目标信息为连接指向的页面内容。如链接标签内容包含“出版社”，则“出版社”的目标信息就是该链接所指的页面。

·若关键字出现在标题中(如包含在〈H1〉〈/H1〉中)，则目标信息是紧跟它后面的直到下一个标题间的字符串，若该标题为文中的最后一个标题，则结束为一个空行或〈HR〉或〈BR〉标记出现。

·若关键字出现在项目(item)或列表中，则目标信息为紧跟它后面的直到下一个〈LI〉或〈UI〉或表尾之间的字符串。

·若关键字是表(table)中的一个域，则对于纵向排列的表来说(列名所在域中不含〈TH〉)，目标信息是关键字所在位置右边的域，而对于横向排列的表来说，则是其下面的域。

3.3.2 *基于模式的数据抽取* 就是用户给定一个模式串，在 www 页面中进行串匹配，根据匹配结果，从中抽取出所需要的值。

所谓模式，是指含有常量和变量的字符串，该字符串包含在一对方括号中，其中变量用/作为开始符号，后跟变量名。如[Mr. /name]，其中“Mr.”为常量，“/name”为变量，整个串即是一个模式。

进行基于模式的数据抽取时，用户指定一个模式后，系统首先在网页中定位模式中的固定单词，匹配成功后，提取出相应的值赋给变量。比如上例中，系统首先定位到“Mr.”，然后将其后的字符串赋给变量“/name”。一般来说，我们主要关心的是名词性短语，因为大多数我们感兴趣的信息多是用名词短语或数值来表示的，而动词经常用来表示行动或状态。

两个或多个模式也可以用布尔操作符进行联结，如：[Mr. /name] or [Ms. /name]，表示同时匹配“Mr.”或“Ms.”开头的文本串，其后的字符串将赋值给变量“/name”。在进行串匹配时，可能会找到超过一个以上的单词可以被赋值给一个变量，这样，我们将计算一个单词的匹配次数，并选择最高出现概率或两个或多个中有同样计数次数中的第一个。

此外，通配符可用于模式中，如：“*”、“+”、“?”等，“*”表示零或多个字符串，“+”表示一或多个字符串，“?”表示零或一个字符串。这样[Dr. /Name received /Degree from * in /Year]就表示匹配任意顺序出现“Dr.”、“received”、“from”、“in”的字符串，找到后，将相应的字符串赋给变量“/Name”、“/Degree”、“/Year”。

3.3.3 *基于样本的数据抽取* 根据用户给定的一个样本来抽取信息，它基于以下假设：一个小范围的 Web 页具有相似的结构和风格。一个典型的例子是一个学院的所有 Web 页均由同一人设计，而这些 Web 页具有相似的结构和风格。因此，当一个使用者想要查询所需信息时(如 email 地址)，他会先手动地从一个 Web 页中定位一个样本，然后告诉系统他想从其他 Web 页获得相似的信息，系统再自动帮他完成。

3.4 基于 XML 的信息抽取

最后再简单介绍一下基于 XML 的信息抽取[13]。当前的 Web 信息大多数都是 HTML 格式的，由于 HTML 具有结构简单性和灵活性，它极大地促进了信息产业的发展，但是也正是由于 HTML 结构太灵活和自由，造成了一个致命的缺陷：难以检索或者抽取隐藏其中的数据。针对 HTML 的这种缺陷，XML 语言应运而生，它一方面继承了 HTML 的灵活性和简单性，另一方面又对其存在的问题做了很大的改进，最重要的就是强制结构的完整性和标签的自定义性。正因为 XML 比 HTML 具有更多的优点，人们普遍认为：XML 最终会取代 HTML 而成为 Web 的通用语言[14]。此外，针对 XML 的研究以及支持 XML 的工具也不断涌现。为此，我们设想：能否将 HTML 格式的文档转换成 XML 格式，然后再进行对 XML 文档进行信息抽取？事实证明这种方法是可行的。具体抽取过程如下：

(1)获取信息源，并将 HTML 转换成 XML. Tidy[15]，可

用于改正 HTML 文档中的常见错误并生成格式编排良好的等价文档，还可以使用 Tidy 来生成 XHTML[16](XML 的子集)格式的这些文档。

(2)找数据中的引用点。无论是在 Web 页面还是 XHTML 视图中的绝大多数信息都与我们完全无关。在这一步中我们的任务就是在 XML 树中找出一个特定区域，并从中抽取我们感兴趣的数据而无需关心外来信息。完成这一任务的最简单的办法通常是，首先检查 Web 页面，只需要看一下页面就可以知道信息位于页面的位置(称之为锚)，然后使用 XSL 来转换我们的 XML，利用 Xpath[17]表达式来指定从根元素到锚的路径。如：/html/body/center/table [6]/tr [2]/td [2]/table [2]/tr/td/table [6]。但是这个方法的缺陷就在于会导致我们对页面布局的修改非常容易遭到破坏。较好的方法是根据周围的内容指定锚。通过这个方法，我们对上例中的 Xpath 表达式重新构造：// table[starts-with(tr/td/font/b,'Appear Temp')]，这样我们就很容易找到表格中粗体显示的 'Appear Temp' 信息了。

(3)将数据映射成 XML。从前一步中得到锚，我们可以创建实际抽取数据的代码。这个代码将以 XSL 文件的形式出现。XSL 文件的目的是标识锚，指定如何从这个位置获取我们正在查找的数据，并且用我们所需的格式构造一个 XML 输出文件。

小结 Web 信息抽取是一个年轻的研究领域，尽管目前该领域研究已经取得了一定的进展，但仍然存在一些问题。首先，信息抽取系统的准确性和健壮性有待提高；其次，在一个新领域上建立信息抽取系统需要许多该领域专家和熟悉 NLP 系统的计算语言学家的共同努力，既费时又费力。随着网络在国内的迅猛发展，Web 信息抽取会变得越来越重要，希望有更多更好的技术能够应用到该领域。

参 考 文 献

1 Cardie C. Empirical methods in information extraction. AI Magazine，1997，18(4)：65～79

2 朱靖波，姚天顺. 中文信息自动抽取. 东北大学学报，1998，19(1)

3 Muslea I. Extraction Patterns for Information Extraction tasks：A survey. In：AAAI-99 Workshop on Machine Learning for Information Extraction，1999

4 朱明. 互联网信息智能搜索和获取方法研究：[中国科学技术大学博士论文]. 2001

5 Soderland S. Learning to extract text-based information from the world wide web. In：Proc. of Third Intl. Conf. on Knowledge Discovery and Data Mining (KDD-97)

6 Huffman S B. Learning information extraction patterns from examples. In Connectionist，Statistical，and Symbolic Approaches to Learning for Natural Language Processing，Springer Verlag，Berlin，1996，1040：246～260

7 Kim J，Moldovan D. Acquisition of Semantic Patterns for Information Extraction from Corpora. In：Proc. of the Ninth IEEE Conf. on Artificial Intelligence for Applications，Los Alamitos，CA：IEEE Computer Society Press，1993. 171～176

8 Califf M，Mooney R. Relational Learning of Pattern-Match Rules for Information Extraction. Working Papers of the ACL 97 Workshop on Natural Language Learning，1997. 9～15

9 Soderland S. Learning information extraction rules for semi-structured and free text. Machine Learning，1999，34(1-3)：233～272

10 Freitag D. Information extraction from html：Application of a general learning approach. In：Proc. of the 15th Conf. on Artificial Intelligence (AAAI-98)，1998. 517～523

11 Muslea I，Minton S，Knoblock C. A hierarchical approach to wrapper induction. In：Proc. of the Third Intl. Conf. on Autonomous Agents(AA-99)

12 Guan T，Wong KF. KPS：a Web Information Mining Algorithm. Computer Networks，Elsevier，1999，31：1495～1507

13 Myllymaki J. Effective Web data extraction with standard xml technologies. WWW10，Hong Kong ACM 1-58113-348-0/01/0005，2001

14 Extensible Markup Language (XML). W3C Recommendation，Feb. 1998. http：//www.w3.org/TR/REC-xml

15 HTML Tidy. http：//www.w3.org/People/Raggett/tidy/

16 XHTML：The Extensible HyperText Markup Language. W3C Recommendation，January 2000. http：// www. w3. org/TR/xhtml1

17 XML Path Language (XPath). W3C Recommendation，November 1999. http：//www.w3.org/TR/xpath.html

(上接第37页)

Semantic DS	Semantic	描述了由多媒体内容模板所描绘或者与之有关的一个叙述性的世界。
Package DS	Package	描述了工具的一种树状组织。
Video Editing DS	(To be submitted)	
Summary Preferences DS	SummaryPreference	详细说明用户对多媒体内容的非线形导航和访问偏好。
Extended Textual Type	ExtentedTextualType	
Phonetic D	Phonetic	描述各种语音信息。

参 考 文 献

1 MPEG-7标准文档 Overview of the Mpeg-7 Standard N4031

2 MPEG-7标准文档 MPEG-7 Context and Objectives N2861

3 MPEG-7标准文档 MPEG-7 Requirements Document N4320

4 MPEG-7标准文档 MPEG-7 Applications Document N3934

5 MPEG-7标准文档 MPEG-7 Projects and Demos N4034

6 MPEG-7标准文档 MPEG-7 Interoperability，Conformance Testing and Profiling Version 2 N4039

7 http：//archive.dstc.edu.au/mpeg7-ddl/ MPEG-7 DDL 主页

8 MPEG-7 eXperimentation Model (XM) Software version 5.5