基于启发式规则的Web信息抽取技术研究

廉成洋[1],毛宇光[1,2],黄玉明[1]
(1.南京航空航天大学 信息科学与技术学院,江苏 南京 210016;
2.南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

摘 要:随着Internet的发展,Web挖掘技术越来越重要,其中的Web信息抽取技术逐渐成为热点,逐渐成为Web挖掘技术的关键技术之一,对Web信息抽取技术的深入研究也为构建更好的面向主题的搜索引擎提供了思路。文中对Web信息抽取的现有技术以及现有技术存在的问题进行了详细的论述。根据Web信息抽取的原理,依据软件工程的观点对Web信息抽取技术提出了具有指导意义的8条启发式规则。在这些规则的指导下,着重阐述了Web信息抽取中的基于结构和内容的信息抽取。通过理论分析及相应的实验说明所提出的8条规则对Web信息抽取具有良好的指导意义。

关键词:Web信息抽取;网页过滤;启发式规则
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)08-0004-04

Web Information Extraction Technology Research Based on Heuristic Rules

LIAN Cheng-yang[1],MAO Yu-guang[1,2],HUANG Yu-ming[1]
(1.College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016,China;
2.State Key Laboratory for Novel Software Technology at Nanjing University,Nanjing 210093,China)

Abstract:Web mining becomes more and more important along with the popularity of Internet. Web information extraction is a hotspot, and becomes one of key technologies in Web mining, and it gives a solution to construct a better theme-oriented search engine. Describes the Web information extraction technologies at present and analyses the problems with these technologies. Then according to the principle of Web information extraction and the principles of software engineering, bring forward eight heuristic rules which can guide Web information extraction significantly. And under instructions of these rules, focus on the structure and content-based information extraction in Web information extraction.

Key words:Web information extraction;Web clear;heuristic rule

0 引 言

随着Internet的飞速发展,Web已经发展成为一个全球的、分布和共享的信息空间。但是目前Web上的数据大部分都是以HTML形式出现的,主要目的是为了显示,缺乏对数据本身的描述,不包含清晰的语义信息。这使得应用程序无法直接解析并利用Web上海量的信息,造成资源极大的浪费。为了增强Web中数据的可用性,出现了Web信息抽取技术,它通过"包装"现有信息源,将网页中的信息以更为结构化、语义更为清晰的方式发布出来,为应用程序利用Web中的数据提供了可能。

文中首先对Web信息抽取的现有抽取技术以及现有抽取技术存在的问题进行了论述。然后根据Web信息抽取的原理,依据软件工程的原理对Web信息抽取技术提出了具有指导意义的8条启发式规则。并在这些规则的指导下,着重阐述了Web信息抽取中的基于结构和内容的信息抽取。

1 Web信息抽取现有技术及存在问题

随着人们对Web信息抽取技术越来越重视,抽取技术随着需求的增加而不断丰富,近年来国内外涌现了多种信息抽取方法,根据抽取原理和抽取方式的不同,分为以下几类:

(1)基于自然语言处理方式;

收稿日期:2008-11-30;**修回日期**:2009-02-20
基金项目:教育部计算机软件新技术国家重点实验室(南京大学)基金(A200711)
作者简介:廉成洋(1985-),男,山东济宁人,硕士研究生,研究方向为数据库系统及应用,Web数据挖掘;毛宇光,副教授,博士后,硕士生导师,研究方向为数据库系统及理论,数据挖掘与数据仓库。

(2)基于包装器归纳方式;

(3)基于 Ontology 方式[1];

(4)基于 HTML 结构方式;

(5)基于视觉的信息抽取。

1.1 基于自然语言理解方式的信息抽取

基于自然语言处理技术的信息抽取是按照人类理解语言的方法对数据进行分析处理的。一般处理步骤包括:句法分析、语义标注、专有对象的识别(如人物,公司)和抽取规则。具体地说就是利用子句结构、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。其中的规则由人工编制,也可从人工标注的语料库中自动学习获得。这类信息抽取主要适用于源文档中含有大量文本的情况。

基于自然语言的信息抽取技术是将 Web 文档视为文本进行处理的。优点是能够基于理解语义进行抽取。缺点很明显:抽取的实现无法利用 Web 文档独特于普通文本的层次特性和标签,抽取规则表达能力有限,缺乏健壮性;获得有效的抽取规则需要大量的样本学习,达到全自动抽取较难,速度较慢,对于操作网上海量数据来说并不适合。目前采用这种原理的典型系统有 RAPIER[2]、SRV[3]、WHISK[4]。

1.2 基于视觉的信息抽取

通常情况下,在分辨语义块的时候,用户会使用一些视觉因素来进行帮助,比如背景颜色、字体颜色和大小、边框、逻辑块和逻辑块之间的间距等等。如果充分地使用 Web 页面的视觉提示信息,并结合 DOM[5]树进行页面语义分块,则可以弥补仅使用 DOM 树所带来的一些缺憾。

微软亚洲研究院提出了 VIPS(Vision - Based Page Segmentation)[6]算法用来抽取给定网页的语义结构。这种语义结构是层次性的结构,在该结构中,每一个结点代表一个语义块。每一个语义块都定义一个 DoC (Degree of Coherence)值来描述该语义块内部内容的关联性。DoC 的值越大,则表明语义块内部的内容之间的联系越紧,反之越松散。

在 VIPS 算法中,DoC 的值位于 1 到 10 之间。不过这个范围是可以更改的。在对 Web 页面进行语义分割之前,首先设定一个预定义 DoC 值 PDoC,通过该值来限定分割的语义块的粗糙程度。当语义块的 DoC 值达到 PDoC 之后,迭代分割就停止。PDoC 越小,则分割的语义块就越粗糙,反之,分割的语义块就越精细。VIPS 算法充分利用了 Web 页面的布局特征:它首先从 DOM 树中抽取出所有的合适的页面块,然后根据这些页面块检测出它们之间的所有的分割条,包括水平和垂直方向。最后基于这些分割条,Web 页面的语义结构将被重新构建,整个 VIPS 算法是自顶向下。

整个过程用图 1 描述。它具有三个步骤:页面块抽取、分隔条抽取以及语义块重构。这三个步骤联合一起作为一次语义块检测的完整步骤。

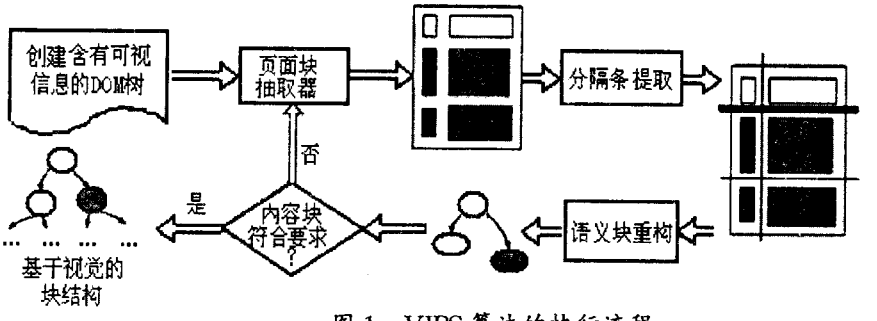

图 1　VIPS 算法的执行流程

1.3 Web 信息抽取存在的问题

通过对现有的 Web 信息抽取技术的分析,发现 Web 信息抽取面临着挑战,这些挑战是有待进一步研究和解决的问题。

(1)抽取规则一直是信息抽取过程的一个重要环节。目前各类信息抽取技术中生成规则的依据主要有五类:结构特征、位置特征、显示特征、语义特征和引用特征。这些方式各有缺陷,如何尽量把上述方法结合在一起共同完成信息抽取是需要解决的一个问题。

(2)机器学习往往通过大量的样本学习来提高获取规则的自动化程度,这意味着系统需要经过较长时间的学习才能获得较好的查准率。抽取规则的适应性较差,缺乏健壮性。如何处理效率与健壮性之间的矛盾是一个重要问题。

(3)性能较好的信息抽取技术需要用户的大量参与,自动化程度不高;而自动化程度高的信息技术其准确率和适应性较低。这两者之间的矛盾也需要解决。

2 Web 信息抽取的启发式规则

传统方法在进行 Web 信息抽取研究时,人们往往只从获得的网页的角度出发按 Web 挖掘的技术来对其研究,然后提出一些复杂的思想和算法。其实,如果基于 Web 网站开发人员和管理维护人员的角度进行思考,基于软件工程的角度进行考虑,很多问题就能找到更佳和更自然的处理方法。

为此,文中从网站开发人员和软件工程角度提出 Web 信息抽取的以下启发式规则:

规则 1　网页(及网页的相关资源)逻辑上若按层

次分类,则在服务器端一定是分层分类存放的。

规则2 同层网页的风格和结构是相似的。

规则3 网页都是由信息块构成的,信息块是逐层嵌套的。

规则4 信息块(以 <table><tr><td>等作为标志)是分析的基本单位。

规则5 网页按DOM树分层,对信息块按层次分析,则每个信息块逻辑上都有其独特作用。可以为每个信息块作分类标记(以表明此信息块的功能或作用,如是否广告、导航条等不相关信息)。

规则6 关键正文信息一定在网页的视觉中心,即:网页中间。

规则7 网页正文部分的信息块内含的标签的种类和数量一定比周围少(正文格式较简单),其正文部分最常见的标签为<h1~hn><p>
。

规则8 页若有注释,则注释信息对信息抽取很有用。

利用上述规则,许多看似复杂的网页抽取算法就能立刻在处理思想上找到依据,同时也能进一步对这些方法的不足提出改进的意见。

其中规则6,在判断块信息是否在网页中间时,可以通过文献[7]页面结构分析技术得以确定,也可通过浏览器简化模拟的方法[8]确定。

利用规则7,只要统计一个信息块的正文标签比和所含标签的种类就可以判断当前信息块是否为正文,在判断时,采用先转化为DOM树,再判断DOM树的每个节点是否是信息块。

3 基于启发式规则对内容抽取的启示

在抽取正文时,有人只是简单地对所有标签和脚本进行过滤,最后只留下一些简单的文本,这样做的后果是:失去了大量标签所含的信息,对后期抽取文章标题、生成摘要等分析处理造成了困难。

有研究者对网页进行抽取时,利用DOM树,但却假设所有正文只存在于table中,这种假设显然只能对部分情况适用。有研究者对网页抽取时,只针对特定的数据源,生成一系列的抽取规则,虽然抽取的精度很高,却不具有通用性。

其实完全可以根据规则1~8,设计出更好的垃圾过滤和通用正文抽取算法。下面根据启发式规则说明一个通用的正文抽取算法。抽取正文的过程可以分为两个步骤:将网页分块和对信息块的取舍。具体包括:网页规范化,建DOM树,分块,为块做标记,删除无用的信息块。

这里最关键的问题是:怎样分块?怎样判断信息块的类型?

3.1 信息块的分类—块特征

块特征主要包括两大方面:空间特征和内容特征。空间特征包括块的位置和大小等等;内容特征包括文字长度,链接数量,图片数量等等。

①空间特征。空间特征主要由4个参数表述:块中心的横坐标(Blockcenterx)、纵坐标(Blockcentery)、块的宽度(Blockwidth)和块的高度(Blockheight)。采用kovacevic中的模拟浏览器显示算法可以计算出相关位置坐标。

②内容特征。内容特征:即一个块表达了什么形式的内容,包括以下几个参数:文字长度(Textlength),链接数量(Linknum),链接文字长度(Linktextlength),图片数量(Imagenum),图片大小(Imagesize)等等。

为了进一步方便分析,现给出块内链接文字和文字长度比LT的定义:设LT值为链接文字长度比上文字长度,见下式:

$$LT = Linktextlength/Textlength$$

通过大量实验和分析得出表1,利用其可以对网页上的各种信息进行分析判断并给出块标记。

表1 判断信息块的方法

块类型		判断各种信息块的方法
链接块		Textlength=0或LT>0.5
文字块		LT<0.5
网页正文块		位置:中间;LT<0.5;内含大量<p>和少量<h1~h6>
噪音信息	广告	必是链接块(Textlength=0,LT>0.5);多为图片链接或文字链接;同层网页一般多页雷同
	导航栏	位置:上方或左侧
	网站目录	一般含有:"相关……","最新……"等敏感词
	版权	位置:最下。倒过来搜索的第一个最大信息块一般含有:"版权所有","copyright","关于……"等敏感词

综合考虑空间特征和内容特征,可以判断出每个块的重要度。最后提取出重要度最高的内容块中的文字,合并这些文字就是所要的正文。

3.2 确定正文信息块的算法

算法总的思想是:以容器标签(table、div等)为依据进行分块。分块时尽量由内到外找到最大的信息块,使此信息块作为一个整体。将网页分块,就要确定DOM树中哪些节点可以成为分块节点。分块节点是指该节点里面的内容可以独立成为一个内容块。分块节点的大小由分块粒度的粗细决定,过粗或过细都将导致抽取结果不完整或保留多余信息。步骤如下:

①判断Web页面采用哪些容器标签。

通过统计DOM树中各种容器标签的数量进行判断,如果某种容器标签被用来做布局,那么它的数量一

定很多。

②考察最底层容器标签节点。

大多数网页一般采用 table 或 div 容器标签来布局,不失一般性以 table 标签为例说明确定正文信息块的算法,div 等其他处理方法类似。

一般正文及有用的信息多数情况下会出现在容器标签嵌套的最里层,因此先从最里层的 table 节点考察起,并对 DOM 标签树采用自底向上的顺序进行处理。即:把标签树中各个最底层的 table 节点下的所有文本节点(TextNode)合并,并统计该块的信息含量。

③判断一个节点能否成为分块节点。这是分块算法的关键步骤。

分块时,一定要尽量往父节点上靠,从而得到逻辑上的最大信息块。因为同一块内的节点标签相似度强,所以在考察时,确定当前节点的块类型,分析兄弟节点的块类型是否一致,若基本一致,则当前节点上移至父节点,继续按上述方法分析,直到发现兄弟节点类型不同时结束。此时的节点就是一个最大信息块。

4　结束语

Web信息抽取的前期研究主要集中在生成和使用包装器阶段,但该方法的缺陷是缺少灵活性和可扩展性;后期研究则基于 HTML 结构,将 Web 文档解析成语法树并对其进行分析[9],通过自动或半自动的方式产生抽取规则,然后进行信息抽取。文中首先对Web信息抽取的现有抽取技术以及现有抽取技术存在的问题进行了详细的论述。然后根据 Web 信息抽取的原理,依据软件工程的原理对 Web 信息抽取技术提出了具有指导意义的 8 条启发式规则,并在这些规则的指导下,着重阐述了 Web 信息抽取中的基于结构和内容的信息抽取。

参考文献:

[1] 陈　静,朱巧明,贡正仙.基于 Ontology 的信息抽取研究综述[J].计算机技术与发展,2007,17(10):84-86.

[2] Califf M, Mooney R. Relational Learning of Pattern-Match Rules for Information Extraction[C]//In Proceedings of the Sixteenth National Conference on Artificial Intelligence. Orlando, Florida: [s. n.], 1999.

[3] Freitag D. Machine Learning for Information Extraction in Informal Domains[D]. [s. l.]: Carnegie Mellon University, 1998.

[4] Soderland S. Learning Information Extraction Rules for Semi-structured and Free Text[J]. Machine Learning, 1999, 34(1-3):233-272.

[5] 李效东,顾毓清.基于 DOM 的 Web 信息提取[J].计算机学报,2002,25(5):526-533.

[6] Deng Cai, Yu Shipeng, Wen Jirong, et al. VIPS: a Vision-based Page Segmentation Algorithm[R]. Microsoft Technical Report, 2003.

[7] 常育红,姜　哲.基于标记树表示方法的页面结构分析[J].计算机工程与应用,2004,40(16):129-132.

[8] Kovacevic M. Recognition of Common Areas In Web Page Using Visual Information: A possible application in a page classification[C]//Proceedings of ICDM02. Maebashi, Japan: IEEE Press, 2002.

[9] 仲　华,崔志明.基于 XML 的信息抽取和多层向量空间技术研究[J].计算机技术与发展,2007,17(7):49-52.

(上接第3页)

(2)比较传统的直接使用关键字来进行信息检索的主流网络信息检索引擎,这里使用的是在本体论基础上经过语义逻辑推理器处理后的语义索引词作为查询的输入,使查询变得准确快速而且节省网络资源。

(3)在索引库的建立过程使用索引器对结果进行语义标引,形成索引库,因为适当的语义知识表示是建立本体库系统的关键,所以良好的语义标引为搜索时利用本体领域奠定基础。

(4) 基于本体的信息检索服务器是功能的核心,智能搜索引擎需要具备符合用户信息需求的本体知识库,在用户界面接受到查询式进行搜索时,搜索器根据已有的本体库,分析查询式,了解检索词的意义,并运用一定的描述语言,根据本体库里概念间的联系产生联想和推理,从而找全找准与用户需求最相关的文档。

除此之外,这个模型中还有一个改进的搜索器专门处理全文检索。

参考文献:

[1] 孙宝传.互联网未来的发展走向综述(I)[J].中国传媒科技,2003(3):61-62.

[2] Berners-Lee T, Handler J, Lassila O. The Semantic Web[J]. The Scientific American, 2001, 284(5):34-43.

[3] 朱礼军,陶　兰,黄　赤.语义万维网的概念、方法及应用[J].计算机工程与应用,2004(3):79-80.

[4] 李选如,何洁月.语义集成:本体映射方法研究[J].计算机技术与发展,2007,17(2):127-130.

[5] 何绍华,宫兆晖.基于语义网的网络信息检索相关性研究[J].情报杂志,2007(2):120-121.

[6] 王文峰,赵　莉.语义网中的本体对象研究及应用[J].枣庄学院学报,2007(2):52-53.

[7] 杨晓青.一种利用 RSF(S)建立本体论的方法[J].计算机应用研究,2002,19(4):11-15.