

一种面向属性的论坛自动抽取方法

刘继勇^{1,2} 曲文龙^{1*}

(石家庄经济学院信息工程学院¹, 石家庄 050031; 同方股份有限公司研发部², 北京 100085)

摘要 基于现有网上论坛信息抽取的不足,提出一种面向属性的论坛自动抽取方法。该方法运用论坛概念模型(Ontology)自动构造包装器,较好地解决了现有的抽取方法准确性差、通用性不强的问题。试验结果表明提出的方法是有效的。

关键词 本体论 包装器 信息抽取

中图法分类号 TP311.52; **文献标志码** A

自网上论坛诞生 20 多年以来,随着互联网技术的发展,已经由原来简单的电子公告板系统发展为功能丰富的网上论坛和虚拟社区模式。几乎每个门户网站都开设了论坛频道,现存在的主要论坛站点有,强国论坛、新华网论坛、天涯论坛、猫扑论坛、凯迪、西祠胡同、新浪论坛、搜狐社区、凤凰论坛等等,网络论坛成为人们探讨问题、交流观点的场所。论坛站点中积存了丰富的信息资源,不但有各类技术资料 and 新闻文档,还包含着用户的判断和评论,论坛站点已成为 Web 信息库的重要组成部分^[1]。

对网络论坛的信息抽取不完全同于一般的信息抽取,主要目的是抽取论坛中用户所需要的内容而不是抽取细粒度的数据。目前在这方面,已经有不少方法被提出,总的来说分为两类,其中一类方法是基于论坛网页位置,具有代表性的是 Lin 和 Ho 提出的 Info Discover 系统,根据 Table 标签把网页分成若干个内容块,然后把词作为特征抽取出来并计算其熵值,再计算每个内容块的熵值,最

后根据熵的阈值来划分内容块的有关性。尽管效率得到了提高,但抽取的信息有很多无关性内容。还有一类方法注重于本体 Ontology 的网页信息抽取,近年来,基于 Ontology 的信息技术受到了大量的关注,并广泛应用于半结构化信息抽取技术中^[1]。本体论是哲学的分支,是研究客观事物存在的本质,它与认识论(Epistemology)是相对的。在自然语言处理中,Ontology 被认为是“特定领域内概念及概念之间关系的集合”^[2],它能有效地表达特定领域内的概念、实体、关系等通用知识,能够帮助提高信息抽取系统的抽取性能。更为重要的是基于 Ontology 的信息抽取模型非常适合作为下一代 Web 技术的通用语义抽取模型,因为下一代 Web 技术—Semantic Web^[3]是基于本体的 Web 技术。但是现在的领域 ontology 基本上是展现出来供标注过程使用的,而无法自动接收标注完的反馈信息。因此,该方法使用的效率低,需要增加机器学习的方法加以完善。

针对以上不足,现提出一种面向属性的论坛自动抽取方法,该方法运用论坛概念模型(Ontology)自动构造包装器,可以是列表页面或详细页面,如图 1 所示。

2009 年 7 月 27 日收到

第一作者简介:刘继勇(1980—),男,硕士研究生,研究方向:数据挖掘与知识发现、搜索引擎。E-mail:sewmdm@hotmail.com。

* 通信作者简介:曲文龙,男,博士,副教授,硕士生导师,研究方向:复杂类型数据挖掘、数据挖掘与知识发现。

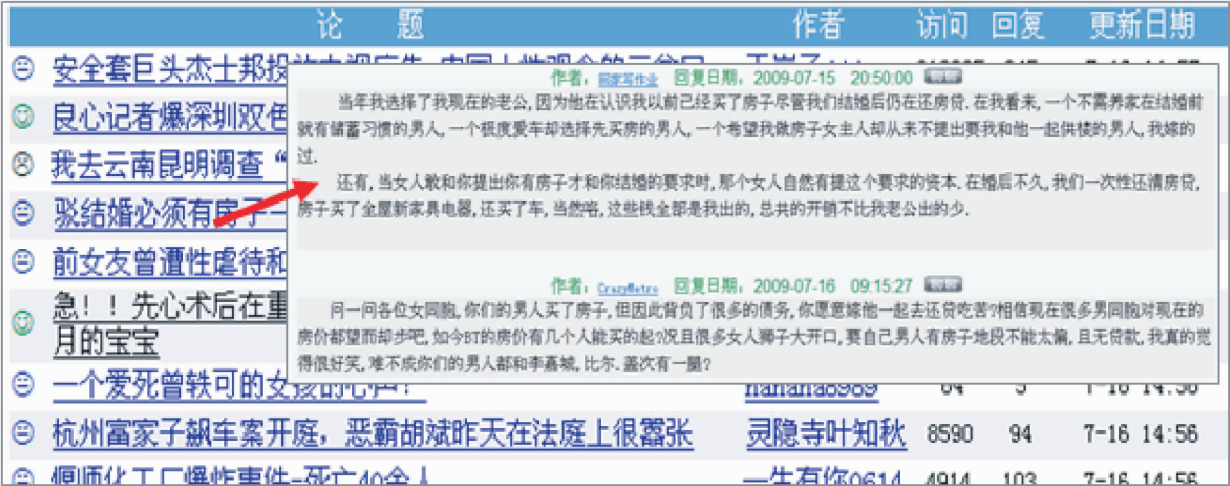


图1 来自天涯社区的列表页面和详细页面举例

1 系统框架

系统的基本框架如图2所示。输入论坛网页，然后对其进行网页的处理，构建包装器，把网页中的主要内容提出，生成数据区域集合，接着对提出的数据进行 Ontology 比对，生成属性数据集合，最后把经过规则系统处理以后的数据存入内容数据库。

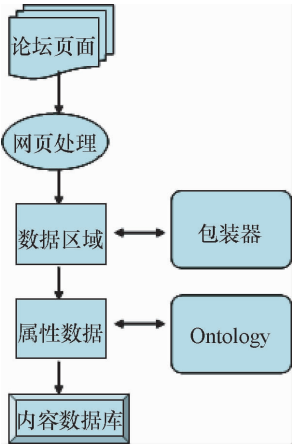


图2 系统框架

1.1 包装器

包装器 (Wrapper)^[7]能解析源文档并将源数据转换为结构化的形式, Wrapper 可以手工编写、半自动化编写或者自动生成。手工生成和半自动生成有两个缺点: ① 需要建立初始化的 Wrapper; ② 当

源文档改变时需要及时更新 Wrapper。

1.2 Ontology

是本系统的基础与核心,构造出良好的面向应用领域的 Ontology 对提高信息抽取的精确度有直接的影响。Ontology 中的概念是基本的元素,所有 Ontology 的理论和实现都是建立在此概念上的。一个领域的概念主要包括涉及这个领域一些实例词、关键词,以及对这些词进行分类而产生的抽象概念,这些关键词,抽象的概念组成了一棵领域概念树。^[8]

2 包装器的产生

通过分析网页结构构造一个符号化的后缀树,再使用重复模式算法进行查找。

2.1 构造符号化的后缀树

后缀树是一种数据结构。一个具有 m 个单词的字符串 S 的后缀树 T , 就是一个包含一个根节点的有向树, 该树恰好带有 m 个叶子^[4,5]。构建长度为 m 的字符串 S 的后缀树, 首先将后缀 $S[1, \dots, m]$ 作为一条单边加入到树中。然后将后缀 $S[1, \dots, m]$ 加入到成长的树中, 其中从 2 增长到 m 。考虑到后缀树中的循环总是以一个头标签为开始, 所以在构造后缀树的过程中仅仅将带有头标签的子串插入到后缀树即可。这样构造的后缀树减小了规模, 也相应的缩减了遍历后缀树的时间, 提高了抽取的

效率。为查找后缀树中重复的模式,需要遍历后缀树和每一个非叶子节点以便检查其所有的孩子节点是否有连续的子串能被发现。例如标记串为

`<body><table><div><td>content1</td><td>content2</td><td>content3</td></div></table></body>`,为标记串生成相对应的后缀树。在后缀树节点下发现的重复个子串,它们都以`<td>`开始并且是连续的。这样,它们就构成了一个连续重复的模式。也就是说被发现的重复串为`<td>(.+?)</td>`,输出到包装器的标记串变为(用正则表达式表示)

```
[^<]*?<body>[^<]*?<table>[^<]*?<div>[^<]*?<td>(.+?)</td>[^<]*?<td>(.+?)</td>[^<]*?<td>(.+?)</td>[^<]*?</div>[^<]*?</table>[^<]*?</body>[^<]*?
```

2.2 重复模式查找算法描述

重复模式查找算法代码如下。

输入:论坛网页源码 S

输出:重复模式标记串 TagS

RepeatPatterFind(S, Sgroup)

{

Input(S) //输入论坛网页源码 S

Extract(S,SL) //提取 S 中的 HTML 标记,并赋值给 SL

For each Si ∈ SL

Match(Si, S, iCount) //在 S 中匹配 Si,并赋值累计数给 iCount

If (iCount = MaxValue)

InsertGroup(Si, Sgroup) //对出现次数最多的 Si 插入 Sgroup 中

End For

MatchTags(Sgroup, S) //匹配最优重复模式标记串

RepeatPatterFind(S, Sgroup) //递归调用并匹配文本之间的文本串是否存在重复模式

}

ConvertRE(Sgroup, TagS) //最后转化 Sgroup 为正则表达式 TagS

以上的算法所找出的重复模式不仅是在标签层次上的发现,还能确定头标签和尾标签之间的文本串是否重复,并找出全部的重复模式。

3 基于 Ontology 的数据抽取

3.1 Ontology 表示

论坛网页的文档典型地包含一些有待抽取的成分,通过分析这些成分的特殊词法词义形态,就能相对准确地抽取出这些成分,如论题、作者、访问数、回帖数、发帖日期、回帖内容等。

这些信息可通过一些对象、属性、约束及术语或词汇来表示,从而构成了论坛概念模型(Ontology)。

3.2 规则系统

规则系统^[9,10]完成由文本到语义描述的转换,规则可以认为是抽取与领域相关的信息的一组原则,按照这个原则选择生成属性数据的信息。定义规则为一个类,它对外有两个接口方法,即 Attribute() 和 Do(),这样可以把同类型规则放到一个规则类的 Attribute 和 Do 中,大大地减少了 if...then 语句的数目,而且对于概念层次结构图总的每个概念都有自己的规则实例集,以便实现规则自动执行机制。

4 实验结果与分析

信息抽取的主要评价指标是召回率(REC)和准确率(PRE),召回率等于系统正确抽取的结果占所有可能正确结果的比例;准确率等于系统正确抽取的结果占有所有抽取结果的比例。为了综合评价抽取引擎的性能,通常还计算召回率和准确率的加权几何平均值,即 F 指数,它的计算公式如下^[6]

$$F = ((\beta^2 + 1)PR) / (\beta^2 P + R)。$$

式中: β 为召回率和准确率的相对权重。 β 等于 1 时,二者同样重要; β 大于 1 时,准确率更重要一些; β 小于 1 时,召回率更重要一些; P 为准确率; R 为召回率。选取国内 6 个著名的论坛进行测试,分别是强国论坛、天涯论坛、猫扑论坛、凯迪、西祠胡同、搜狐社区。当取 $\beta = 1$ 时对每个论坛抽取 50,500 数量的话题进行测试,如表 1 所示。

表 1 6 个论坛的自动抽取测试结果

论坛	抽取话题数量	可能正确结果	所有抽取结果	正确抽取结果	P (准确率)/%	R (召回率)/%	F (指数)
强国论坛	50	43	63	43	68.26	100.00	0.811 4
	500	560	459	442	96.30	98.21	0.972 5
天涯论坛	50	38	37	37	100.00	97.37	0.986 6
	500	426	430	422	98.14	99.06	0.986 0
猫扑论坛	50	58	57	53	92.98	91.38	0.921 7
	500	487	490	476	97.14	97.74	0.974 3
凯迪	50	47	56	47	83.93	100.00	0.912 6
	500	523	506	478	94.47	91.40	0.929 1
西祠胡同	50	58	48	39	81.25	67.24	0.735 8
	500	508	482	470	97.51	92.52	0.949 4
搜狐社区	50	49	39	38	97.44	77.55	0.863 6
	500	481	455	430	94.51	89.39	0.918 7

从表 1 可知,无论是召回率还是准确率都能够达到一个较高的水平,提出的抽取方法能正确地完成抽取任务。

5 结论

提出一种面向属性的论坛自动抽取方法,该方法运用论坛概念模型(Ontology) 自动构造包装器,较好地解决了现有的抽取方法准确性差、通用性不强的问题,试验结果表明提出的方法是有效的,进

一步的工作就是基于 JavaScript 的论坛自动抽取。

参 考 文 献

1 Fensel D, Harmelen V F, Horrocks I, *et al.* OL: an ontology infrastructure for the semantic Web. *Intelligent Systems*, 2001; 16 (2) : 38—44

2 Neches R, Fikes R E, Gruber T R, *et al.* Enabling technology of knowledge sharing. *AI Magazine*, 1991; 12 (3) : 35—36

3 Semantic Web[EB/OL]. <http://www.w3.org/2001/sw/>. 2006-01

4 Crescenzi V, Mecca G, Merialdo P. Road-runner: towards automatic data extraction from large web sites. Roma; *Proceedings of the 27th International Conference on Very Large Data Bases*, 2001; 109—118

5 Wang Jiyiing, Lochovsky F. Data extraction and label assignment for web databases. *Proceedings of the 12th International Conference on World Wide Web*. New York: ACM Press, 2003; 187—196

6 李保利,陈玉忠,俞士汶. 信息抽取研究综述. 北京: 北京大学计算机科学与技术系计算语言研究所, 2003

7 Marcus M, Santorini B, Marcinkiewicz M. Building a large annotated corpus of english“the Penn Tree Bank” in the distributed Penn Tree Bank Project CD2ROM. *Linguistic Data Consortium University of Pennsylvania*, 2006

8 Maynard D, Tablan V, Cunningham H. Architectural elements of language engineering robustness. *Journal of Natural Language Engineering Special Issue on Robust*, 2005; 12 (7) : 102—108

9 欧健文,董守斌,蔡 斌. 模板化网页主题信息的提取方法. *清华大学学报: 自然科学版*, 2005; 45 (S1) : 1743—1747

10 赵欣欣,索红光,刘玉树. 基于标记窗的网页正文信息提取方法. *计算机应用研究*, 2007; 24 (3) : 144—145

BBS Information Auto Extraction Based on the Attribute

LIU Ji-yong^{1,2}, QU Wen-long^{1*}

(Information Engineering University, Shijiazhuang University of Economics¹, Shijiazhuang 050013, P. R. China;
R&D Center, Tsinghua Tongfang CO. LTD², Beijing 100085, P. R. China)

[Abstract] Based on the shortage of BBS information extraction, a new attribute-oriented method of automatic extraction is presented, this method constructs wrapper automatically in the use of BBS conception model(Ontology). It has reserved the issues of the poor accuracy and not strong commonality of the existing methods of extraction. Test results showed that the proposed method is effective.

[Key words] ontology wrapper information extracting