基于正则表达式的企业主页信息抽取①

靳小川[1]　刘万军[1]　赵　雷[2]　(1.辽宁工程技术大学 软件学院 辽宁 葫芦岛 125105；2. 沈阳师范大学 计算机与数学基础教学部 辽宁 沈阳 110034)

摘　要： 主要分析了企业主页上描述企业基本信息表达语句的结构特点，提出了基于正则表达式的企业主页信息抽取的方法和技术，并设计开发了一个相应的原型系统对一些企业信息项进行抽取。实验结果表明，该系统可以有效地从企业主页上抽取企业相关信息，并得到较高的抽全率和抽准率。

关键词： 企业主页；正则表达式；信息抽取

Enterprise Homepage Information Extraction Based on Regular Expression

JIN Xiao-Chuan[1], LIU Wang-Jun[1], ZHAO Lei[2]

(1. Software College, Liaoning Technical University, Huludao 125105,China;

2. Computer and Math College, Shenyang Normal University, Shenyang 110034, China)

Abstract: The paper mainly analyses the structural characteristic of the sentences that describe enterprise basic information on enterprise homepage. It proposes the method and technique of enterprise homepage information extraction based on regular expression, and develops an archetype system to extract some enterprise information items. The experimental results show that it can extract enterprise-related information from enterprise homepage effectively and get a high recall and precision.

Keywords: enterprise homepage; regular expression; information extraction

1　引言

随着互联网的发展和电子商务的兴起，绝大多数的企业已经建有自己的网站。在这些企业网站中，通常会在企业主页(有的是在企业网站首页，有的是在一个称作“公司介绍”或“公司简介”的专门网页等等，我们这里统一称之为“企业主页”)上进行企业自身的介绍，包括企业名称、企业产品信息、人员信息、企业的联系方式等。如何自动获取这些企业主页，并从中抽取出各自对应的企业相关属性信息，进而发现潜在的商业伙伴或本企业的竞争对手的相关信息及其现状，对企业的生存和发展具有重要的意义。

2　相关研究

在互联网信息抽取领域，已经开展了大量的研究工作。从处理的对象来分，一类是针对整个互联网抽取出所感兴趣的信息。开始于 60 年代中期并一直延续到 80 年代的美国纽约大学 Linguistic String 项目[1]及以后一系列的 MUC（Message Understanding Conference）会议在这一方面做了大量有效工作，主要使用模板和槽以及相应的抽取规则来完成。另一类是针对某单一 Web 页面，剔除与主题信息不相关内容，得到该页的主要信息。美国哥伦比亚大学的 PSL 实验室开发的 Crunch 系统[2]，在这方面做了一些有效的工作。

然而国内外对在互联网上进行企业相关信息主题进行抽取的研究还很少，文献[3]设计并实现了一个对企业相关属性实体信息进行抽取的系统-CAIES，该系统根据中文企业网页对不同企业属性描述的特征采用不同策略来对这些企业属性进行识别和抽取；还有一部分研究主要应用在基于 web 的企业竞争情报收集方面，如文献[4]，这方面研究大多针对性不强。

作者曾在硕士论文[5]中针对如何自动获取企业

① 收稿时间:2009-11-13;收到修改稿时间:2009-12-20

主页并对其进行信息抽取展开研究，其总体思路如图 1 所示。

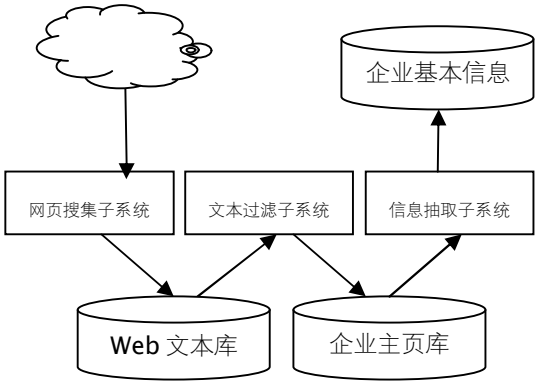

图 1 企业主页信息抽取总体思路

各子系统主要功能说明如下：

网页搜集子系统：根据关键字从 Internet 上搜索网页，并将搜索到的网页下载到本地 web 文本库中。

文本过滤子系统：对 web 文本库的网页进行文本过滤，主要目的是将企业主页筛选出来然后保存到企业主页库中。

信息抽取子系统：对企业主页库的每个网页进行信息抽取，主要目的是将企业基本信息抽取出来，最后保存到企业基本信息库中。

这里主要应用了基于元搜索的网页搜集和基本样本公共特征的企业主页过滤获取企业主页，具体的研究方法这里不再赘述，本文只研究如何在企业主页上抽取企业相关属性信息。

3 基于正则表达式的企业主页信息抽取

对于企业主页的信息抽取,其信息抽取过主要包括三个步骤：获取网页源码、预处理和基于正则表达式的信息抽取。获取网页源码和预处理主要是对企业网站主页的源码进行分析、去除网页标记和网页噪音，从而得到规范的企业介绍文本，这里不作为主要内容介绍，下面主要介绍基于正则表达式的企业主页的信息抽取。

3.1 企业主页的结构特点

在企业主页中，通常含有企业名称、规模、人员等这些企业基本信息，这些信息通常包含某些特殊关键词的语句当中，下面我们针对企业名称、成立年份、企业面积、资产信息、人员信息、生产能力、质量认证、联系方式八个方面的内容，逐个分析这些企业基本信息的语句结构特征。

(1)企业名称

在企业主页中，企业的完整名称通常分布在企业主页的标题中或公司介绍的开头部分。企业名称通常以“有限公司”、“有限责任公司”、“集团”等为句尾，前面通常有“欢迎光临”字样，这些语句的结构特如下：

“<欢迎光临>***（有限责任公司/有限公司/集团）”

其中<欢迎光临>表示有可能包含“欢迎光临”字样（也可以不包含），***表示任意一个或多个文字或数字，（有限责任公司/有限公司/集团）表示包含“有限责任公司”、“有限公司”、“集团”三个词语中任意一个。其他信息项语句结构特征的表达方式采取类似的方式。

(2)成立年份

成立年份一般在表达方式类似为“始建于 1998 年”或“于一九九八年成立”等这样的语句中，常有“成立”、“始建”、“建于”、“年”等这些关键词。这些语句的结构特如下：

“（成立/始建/建于）***年”或“***年***（成立/始建/建于）”

(3)企业面积

企业面积一般在表达方式类似为“占地 1800 余亩”或“面积三十多万平方米”等这样的语句中，常有“占地”、“面积”、“米”、“亩”、“公顷”等这些关键词。这些语句的结构特如下：

“（占地/面积）***（亩/米/公顷）”

(4)资产信息

资产信息一般在表达方式类似为“资产十亿”或“注册资金 10 亿”等这样的语句中，常有“资产”、“注册资金”、“元”等这些关键词，这些语句的结构特征如下：

“（资本/资产/注册资金）***元”

(5)人员信息

人员信息一般在表达方式类似为“拥有职工 60 人”等这样的语句中，常有“职工”、“员工”、“职员”、“人”等关键词，这些语句的结构特征如下：

“（职工/员工/职员）***人”

(6)生产能力

企业的生产能力的表达方式是多种多样的，描述生产能力的语句中常含有的关键词有“生产能力”、“年

产量”、“年产值”等，这些语句的结构特征如下

“***（年产量/年产值/生产能力）***”

(7)质量认证

一个企业可能有被社会公认的质量体系标准，也可能没有。质量体系标准的常用关键词有“ISO”、“认证”等，其语句的结构特征如下：

“***认证”

(8)联系方式

企业联系方式语句的关键字不外乎有“地址”、“电话”、“Email”、“传真”这么几类，其语句的结构特征如下：

“(地址/电话/传真/Email）***”

3.2 基于正则表达式的信息抽取

正则表达式由美国数学家 Stephen Kleene 研究“神经网事件的表示法”时引入，用于描述正则集的代数表达式[6]。它是一串特殊的字符，可根据一定的算法来匹配文本。现在正则表达式主要用于基于文本的编辑和搜索工具，实现数据有效性验证、文本替换以及根据模式匹配从字符串中提取子字符串等。一个正则表达式通常由若干普通字符(字符a到z)以及特殊字符(元字符)组成。如正则表达式“ab*”将匹配有一个 a 后面跟着零个或若干个 b 的字符串(例如"a", "ab", "abbb"等等)。

表 1　企业基本信息项描述语句对应正则表达式

企业基本信息	正则表达式
企业名称	[^(，|。|、|，|.|；|：|光临|)]* (有限公司|有限责任公司|集团)
成立年份	(成立|始建|建)[^(，|。|、|，|.|;|：)]*年[^(，|。|、|，|.|;|：)]*[^(，|。|、|，|.|;|：)]*年[^(，|。|、|，|.|;|：)]*(成立|始建|建)[^(，|。|、|，|.|;|：)]*
企业面积	[^(，|。|、|,|.|;|：)]*(占地|面积)[^(，|。|、|,|.|;|：)]*(亩|米|公顷)
资产信息	[^(，|。|、|,|;|：)]*(资产|注册资金|资本)[^(，|。|、|,|;|：)] *元
人员信息	[^(，|。|、|,|.|;|：)]*(职工|员工|职员)+[^(，|。|、|,|.|;|：)]*人
生产能力	[^(，|。|、|,|.|;|：)]*(年产值|年产量|生产能力)[^(，|。|、|,|.|;|：)]*
质量认证	[^(，|。|,|.|;|：)]*认证
联系方式	(地址|电话|传真|email|Email|E-mail)[^(，|。|||、|,|.|;|：)]*

通过第 2.1 节对企业网站主页上的八个企业基本信息项语句结构特征进行分析，提取出它们的通用表达模式，并将其用正则表达式描述，从而实现对相应信息的自动获取。表 1 为作者总结的八个企业基本信息项描述语句对应的正则表达式。

3. 系统实现与实验验证

为对本论文所提思路进行一个验证，作者应用 Microsoft Visual C# 2005 开发了一个企业主页信息抽取原型系统，如图 2 所示。其核心思路为：使用 C#2005 的 WebBrowser 对象(C# 2005 的前面版本是一个 COM 组件[7])浏览企业主页，并用其属性 DocumentText 取得网页内容字符串,然后用正则式匹配抽取相关内容。在.NET 2005 中，提供了操作正则表达式的相关类在名字空间 System.Text.RegularExpressions 下，其中常用的类有：Regex 类(创建正则式)和 MatchCollection 类(匹配项集合)。类 Regex 的常用方法有：Match(匹配某个正则式)、Replace(替换匹配正则式的字符串)等。

图 2　企业主页信息抽取原型系统

作者通过原型系统对采集的 30 个企业主页进行信息抽取实验，由于每个企业主页中含有的信息量不同和系统的抽取效果原因，部分企业信息项没有抽取结果，例如图 2 中的企业主页没有企业联系方式的文字描述，因此没有抽取出“联系方式”信息。

表 2 是对 30 家企业主页进行抽取后形成的 30 条记录的企业信息项的统计结果。信息项总数 M 是人工对企业主页上的各信息项进行识别和校验的数目，抽取信息项数 N 是原型系统对企业主页抽取的各信息

项数目，抽取正确项数 K 是原型系统抽取结果中经人工校对认为正确的项数。抽全率 R(Recall)描述的是对各属性的抽取全面程度，是正确抽取结果信息项数与人工识别的信息项总数的比值，计算公式是 R=K/M。抽准率 P(Precision)描述的是对各属性的准确抽取程度，是正确抽取结果信息项数与系统抽取信息项总数的比值，计算公式是 P=K/N。

表 2 企业信息抽取统计结果

	企业名称	成立年份	企业面积	资产信息	人员信息	生产能力	质量认证	联系方式
信息项总数M	30	22	19	24	25	22	26	20
抽取信息项数N	26	20	17	18	19	18	24	16
抽取正确项数K	23	17	14	15	15	13	21	14
抽全率R（%）	76.7	77.3	73.7	62.5	60	59.1	80.8	70
抽准率P（%）	88.5	85	82.4	83.3	78.9	72.2	87.5	87.5

从表 2 可以看出，对于某些企业基本信息项，如企业名称、成立年份、质量认证、联系方式的识别效果较好，原因在于这些信息在企业主页上的表达具有鲜明的特征，易于用一定规则进行识别。而其他企业基本信息项也有较高的抽全率和抽准率，说明本系统在企业主页上抽取企业相关信息还是比较有效的。

4 结语

本文通过分析企业主页上对不同企业属性描述语句的结构特征，提出了基于正则表达式的企业主页信息抽取的方法和技术，并设计实现了相应的原型系统对企业名称、成立年份、企业面积、资产信息、人员信息、生产能力、质量认证和联系方式八个企业信息项进行抽取。经实验和测试结果表明，该系统具有较高的抽全率和抽准率，从而验证作者所提思想方法的可行性与准确性。然而，本文的工作还处于一个方法探索阶段，对于各企业属性信息准确抽取还需要深入分析它们的表达特征，而且信息抽取结果只停留在语句层面，还没有达到信息的精确抽取(抽取结果具体到某一个词或一个数字等)，这些都是下一步需要改进的方向。

参考文献

1 李保利,陈玉忠,俞士汶.信息抽取研究综述.计算机工程与应用,2003,39(10):1－5.

2 Gupta S, Kaiser G, Neistadt D.DOM-based Content Extraction of HTML Documents.12th International World Wide Web Conference, May 2003.

3 张丙奇,姜吉发.企业相关信息抽取技术研究与系统实现.微电子学与计算机, 2004,21(1):1－6.

4 陈龙.基于 WEB 信息抽取的企业竞争情报系统研究.合肥:合肥工业大学, 2007.

5 那宝贵.面向合作伙伴选择的中文 WEB 信息获取系统研究.葫芦岛:辽宁工程技术大学, 2007.

6 邱清盈,郑国民,冯培恩,武建伟.基于正则表达式的专利信息提取方法研究.中国机械工程, 2007,18(19):2326－2329.

7 吴伟,刘友华.基于 DOM 的 Web 信息自动抽取.现代图书情报技术, 2004,(2):68－71.