基于规则和非规则方法的 WEB 信息提取

黄晓宏[+]　　连理[*]　　夏迎炬[*]　　徐国伟[+]

[+] 富士通研究开发中心有限公司　　[*] 复旦大学计算机系

[+]{jashuang, guowei }@frdc-fujitsu.com.cn　　[*]{leelix@yahoo.com，xia_yj@263.net}

摘要：互联网上的各个信息源是相互独立的。如果一个系统能够把关于某个主题的来自各个信息源的信息集成到一个完全的信息源中，用户就能方便地获得这个主题他（她）所需的最想要的或者全部的信息。该系统中最重要的一个部分就是从网页中提取指定的信息。本文以网上书店为例详细介绍了 web 页面信息提取的实现。网页中一些信息可以采用基于正则表达式的规则提取，然而也有一类信息很难用规则提取，例如书名。对这些信息，我们采用了基于字体、距离等非语言启发信息的非规则方法提取，试验获得了比较好的结果。以网上书店为例，采用非规则的方法使书名提取的 F 值提高了 31 个百分点。

关键词：　web 信息集成，　web 信息提取，规则，非规则方法

Web Information Extraction Based on Rule and Non_Rule Methods

Huang Xiaohong[+]　　Lian Li[*]　　Xia Yingju[*]　　Xu　Guowei[+]

[+]Fujitsu R&D Center CO., LTD.　　[*] Computer Department, Fudan University

Abstract : Various information source over the internet is independent to each other. If a system could integrate all the information from various sources on one topic into a complete information source, a user would be able to conveniently get the most desirable or all the information on the topic. One of the most important parts of such a system is to extract target information from web pages. Taken online bookstores as a testbed, this paper discusses in detail about implementation of information extraction from web pages. Some kinds of information in the pages can be extracted according to regular expression rules, but there is a kind of information which cannot be extracted according to rules easily, e.g. book names. As for this kind of information a non_rule method for extraction is adopted based on non_linguistics heuristic information such as *font*, distance, and the experiment result is good. The adoption of non_rule method increases the F_measure of book name extraction 31 percent point.

Key words: web information integration, web information extraction, rule, non_rule method

1　引言

World-Wide Web 的飞速发展使它成为信息发布、传播的主要载体。但是提供信息的各

个信息源是相互独立的，彼此之间并不存在必然联系，无论是形式还是内容都不存在一致性。一般来说，我们在 WWW 发现的信息源对他们要覆盖的域来说并不一定是完全的。例如：一个书目不大可能完全包括计算机科学（Computer Science）的文献。只是在某些情况下，我们能够肯定信息源的完全性。例如：DB&LP Database 就包括了大多数主要数据库会议的论文。既然在绝大多数情况下，一个信息源是不完全的，自然希望有一个系统能够把来自各个信息源的数据集成到一个完全的信息源中，这样用户有可能获得他（她）所需的最想要的或者全部的信息。

目前的搜索引擎在查找相关文档方面已经做得不错了，例如 Google，通常情况下提供给用户的文档是相关的，而且速度也很快。但是一般的搜索引擎是基于整个文档的，不是基于细节的，基本上不对页面内容进行深加工，它不能直接回答用户的问题。为了某个特定的信息，用户不得不在不同的页面之间浏览以获得所需的信息。如果用户能够从一个统一的接口通过查询就能获得所需的信息无疑会大大降低工作强度和减少工作时间。在这样一个统一的接口后面则是把从各个站点搜集来的相关数据集成到一个统一的数据库里，这样的技术称之为 Web 信息集成或数据集成。用户获取信息方式的不同是信息集成系统和一般的搜索引擎的的区别所在。

要建立一个 web 信息集成系统，其中最重要的一个部分就是从网页中提取指定的信息，并把这样的信息放到数据库中。

在这样的背景下，本文以网上书店为例详细介绍了 web 信息提取的实现，主要包括三项，HTML 标记分析、规则提取、非规则提取，提供了试验结果，并对其进行了分析，然后介绍了提取后的一致化工作，最后介绍了国内外的相关工作，并对未来的工作进行了展望。

2 WEB 信息提取

2.1 信息提取的信息源

这里处理的 web 页面是网上书店的关于所卖书的具体介绍。这里认为一个页面只包含一本书的数据，也就是一条纪录（这在绝大多数情况下是成立的）。图 1 是一个页面的例子，来源于当当网上书店（中国最大的网上书店）。

我们选取了一般用户最关心的 7 项数据用于提取，分别是书名、作者、出版社、出版时间、定价、现价、分类。

2.2 提取过程

总的来说，提取可以分为两类，一类是可以用规则提取的部分，另一类则是很难用规则提取的，或者即使用规则提取，效率也很低。当然具体到每一个槽而言，不一定完全属于第一类或者第二类，对这样的的槽，则要综合考虑。

基于这样的情况，web 信息提取的步骤如下：

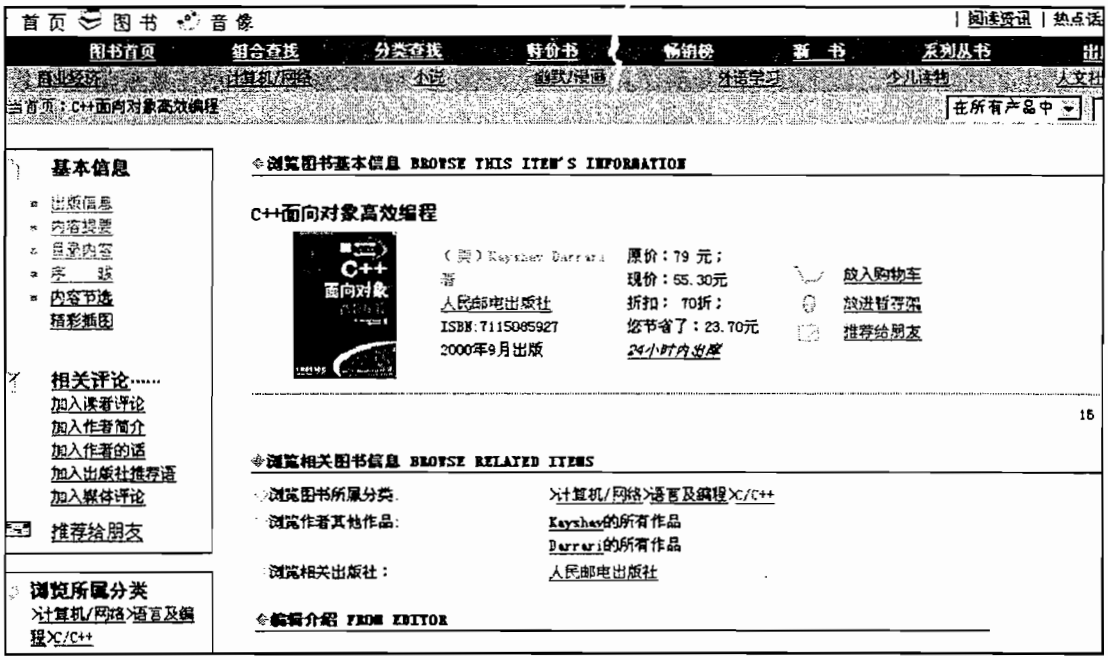

图 1　　书籍介绍的一个页面

(1)　标记分析：对 web 页面的源码进行标记处理，把标记和文本分开，这样 web 页面被分解成两个流，一个是标记流，一个是文本流；

(2)　规则提取：对文本流尝试每一条规则，提取出可能的槽值，并纪录其位置；

(3)　非规则提取：根据已经分析出的结果，用规则以外的方法如标记分析来提取槽值，进一步提高精度。

下面分别介绍这三个步骤，

（1）　对 **web** 页面标记的分析

因为这里主要对文本进行分析，所以这一步的主要任务是找出散落在 web 页面中的各个文本片断。为了以后分析的需要，记录了每个文本片断在原来页面中的位置。按照 HTML 标记的规范，这一步是比较简单的。

需要注意的两个问题是：（1）要把在意义上不可分割的几个文本片断合并在一起，虽然它在 HTML 源码中是分布在不同行上的；（2）需要对分布在不同行上的一个 HTML 标记的全部描述合并在一起用于后面的分析。

为了后面分析的需要，我们还把源码中相邻的、意义上紧密关联的文本片断合并在一起，这主要是针对价格的。我们把“￥”和数字、数字和“元”合并成一个字符串。

（2）　规则提取

在要提取的槽值中，大部分都是有规律可循的。规则的来源可以是人工编写的，也可以

由机器通过学习标注语料获得。在目前的技术水平下，人工规则仍然具有最高的精度，出于实用的角度和初始研究的考虑，我们采用了人工编写规则的方法。

在有了规则集以后，就可以由机器自动提取所需的槽值。

首先介绍一下规则的组成。对应于每一个槽，有一个以上规则，优先级高的规则放在前面。每个规则又分为三个部分，每一部分用正则表达式表示：

（a）前项：预示着对应着目标槽的一个值即将出现；

（b）槽值：要提取的值，槽值本身可能也有前缀或后缀，这些也可以作为规则的一部分；

（c）后项：表明前面出现的字符串是目标槽的一个值。

下面是几个规则的例子，由于这个系统目前是在 Win32 的 ActivePerl 5.6 上实现的，所以这里的规则是基于 Perl 支持的正则表达式的。为了实现的方便，这里把槽值和后项合并在一起。（了解规则之前，不妨熟悉一下一些常用的模式的含义[1]。）

如下所示是对应作者这个槽的两条规则：

规则 1：	覆盖带有前项的情况，可以带有后项，也可以不带后项。
前项：	((\s*作\s*者| ^编著者) (:| ：|\s+|$))
槽值和后项：	(?:^(\D{4,20}?) (?:等)? (?:编著|主编|著|编)) | (\D{4,20}?) (?:等)?译$ | ^\s*(\D{4,20})\s*$
覆盖的例子：	“作　者：王世普主编”、“作　者：邱仲潘等译”、“作　者：晶辰工作室”。
规则 2：	覆盖只有后项的情况
前项：	.*　（代表任意）
槽值和后项：	^(\D{4,20}?)(?:等)? \s* (?:编著|主编|著|编) (?:\D{4,16}?)(?:等)?译$ | ^(\D{4,20}?) (?:等)? \s* (?:编著|主编|著|编) $ | ^(\D{4,16}?) (?:等)? 译$
覆盖的例子：	“（美） 史蒂文斯 著 胡谷雨 等译”

（3）　非规则提取

由于这里要提取的数据项大部分都可以通过规则的方法提取出来，这就为非规则提取提供了良好的上下文基础。

可以认为，我们要提取的数据是在一个局部范围内的。在这样一个范围内，具有某个槽的槽值的某些特征的字符串可以认为是所要提取的值。这里讨论书名的识别。

识别书名是基于这样两点认识：

---

[1] \s　代表空白字符

^　代表从字符串的开始进行匹配

$　代表从字符串的尾进行后退匹配

\D{4,20}　代表匹配非数字字符 4～20 次

（1） 它和书的其他属性在一个共同的比较小的范围之内；

（2） 书名一般要比周围的其他属性显示要显著一些，具体表现在字体大，加黑等。

从这两点出发，书名提取的算法如下所示：

（1） 对已经识别出来槽值，比较其在源文件中的位置，得出其最小的行号，最大的行号；

（2） 认为书名在最小行号向前 Prescope 行和最大的行号向后 Postscope 行的范围内，（在实际操作中 Prescope=10，Postscope=0）；

（3） 在这个范围内对所有文本片断分析限定文本片断的重要标记，根据标记特性对文本片断进行分类，因为书名受比较独特的标记限制，所以丢掉最一般的文本片断；

（4） 在这个范围内，对余下的文本片断分析评估它们的 HTML 标记特性、文本片断本身的特性、距离特性，如 FONT，如果它的大小大于某个绝对值，或比默认字体大一些，或者属于某个特定的字体类，赋予一定正的权值；

（5） 经过以上的步骤，综合评估每个文本片断，值最优的是所要的书名。

同样的方法可以适用于作者的识别，所不同的是，作者具有人名的特点，因此可以认为在一定范围内，具有人名潜在性的文本片断，而且又没有被认为是别的槽值，那么它就是作者这个槽的值。

2.3 试验结果

在我们的试验中，一共选取了 6 个网站的 100 个页面。在测试前，人工标注了这 100 页，对所有应该标注的数据项进行了标记。然后用一个评测工具对机器提取结果进行了评测。表 1 是测试的结果，其中：

$$召回率 = \frac{正确结果}{答案} * 100\% \qquad 精确率 = \frac{正确结果}{提取结果} * 100\%$$

$$F = \frac{2 * 召回率 * 精确率}{召回率 + 精确率} * 100\%$$

槽名	答案	提取结果	正确结果	精确率	召回率	F 值
书名	92	92	87	94.57%	94.57%	94.57%
作者	92	87	77	88.51%	83.70%	86.04%
出版社	95	99	94	94.95%	98.95%	96.91%
出版时间	92	99	91	91.92%	98.91%	95.29%
原价	95	99	94	94.95%	98.95%	96.91%
现价	68	78	68	87.18%	100%	93.15%
类别	39	10	10	100%	25.64%	40.82%
全部	573	564	521	92.38%	90.92%	91.64%

表 1 信息提取结果

表 2 给出了采用非规则提取方法前后的效果对比，可以看到效果还是显著的。这也充分说明了 HTML 标记对 WEB 页面信息提取的重要作用。当然目前对标记的分析还是简单的，如

果做得很复杂，甚至可以利用文本片断在页面中的位置来判断。

书名的提取	答案	提取结果	正确结果	精确率	召回率	F 值
采用非规则提取前	92	48	44	91.67%	47.83%	62.86%
采用非规则提取后	92	92	87	94.57%	94.57%	94.57%

表 2 给出了采用非规则提取提取前后的结果对比

总的来说，结果还是令人满意的，这主要是因为：（1）所提取的页面都是单纪录的；（2）目前所测试的页面的大部分还是有规则可循的；（3）其中一些比较难以提取的项，又可以根据已经顺利提取的项划定的一个局部范围和对 HTML 标记的分析提取出来。

我们还对网上购物的其他领域进行了试验，发现上面的方法是基本有效的。

目前我们还没有对多纪录的甚至嵌套的情况进行试验，这种情况的难度大一些。但这里介绍的方法是同样可以适用的，但要作相应的修改。

2.4 提取信息的一致化

在从多个信息源提取数据时，要遇到一个的问题是非一致性。非一致性可以粗略地分为两类，一类是格式的不一致，一类是文字的不一致。下面分别介绍：

格式的不一致是指一个概念可能有不同的表达方式。通常情况下各个不同的网站的时间、价格表达方式是不一致的。例如同一个时间可能有以下表达方式：2000 年 9 月，2000/09，2000.9；价格则可能有：24.7 元、￥24.7、24.7 等。格式的一致化是比较容易实现的。

文字的不一致是指两个数据源里提及的对象是现实世界同样的一个实体（entity），但由于各个数据源都采取自己的命名习惯和速记法（shorthands）而导致语言文字表现形的式不同。例如：IBM 和国际商业电器公司是指同一个实体,但可能分别被不同的数据源采用。要解决这个问题依赖于本体论（Ontology）。大多数系统用领域特定的启发信息来解决这个问题。这里不做详述。

为了以后信息处理的需要，必须把这些数据的表达方式统一起来，称之为一致化。在我们的工作中，主要对日期、价格、分类做了处理。

3 相关工作

信息提取最早来源于 MUC（Message Understanding Conference），而最早期手工编写的基于 Web 的信息提取程序（wrapper）的一个系统是 TSIMMIS 系统（[Chawathe 94]）。这之后，各种各样的 wrapper 构造系统诞生了。

一些研究工作已经考虑建立工具来快速创建 wrapper。一类工具基于开发专门的语法来确定数据在 HTML 页面是如何排列的，因而确定如何提取数据。第二类技术基于开发归纳学习的技术来自动学习一个 wrapper。通过这些算法，我们给系统提供一个 HTML 页面集，页面里的数据都做了标记。算法使用带标记的例子来自动输出语法，根据该语法能从随后的

页面里提取出数据。把 Wrapper 的构造形式化为归纳学习以及相关的一系列学习简单 Wrapper 类的算法首先由[Kushmerick 97] 给出。[Ashish97]为了获得更快的学习，利用了通常使用 HTML 的特定启发信息。 [Craven 98] 首先试图填补机器学习方法和自然语言处理方法处理 wrapper 构造问题的之间的空隙。

这里应该指出，目前机器学习的方法和人工方法的效率还是有显著差距的。从实用的角度来看，人工方法仍然是首选。当然机器学习的方法代表将来的方向。

4 结束语

本文以网上书店为例详细介绍了 web 信息提取的实现过程，包括 HTML 标记分析、规则提取、非规则提取、数据的一致化工作。应该说，目前的结果是令人满意的，对特定领域的实用是足够的，提出的非规则的方法也是有借鉴意义的。

一些技术的发展趋势将会影响 web 信息提取和集成技术。第一个当然是 XML。XML 格式的数据将不再需要信息提取程序把人可读的数据转化为机器可读的形式。即使这样，从各个 web 源来的数据的语义集成仍将是一个难题。另一个是隐式 web（hidden web）的增长。最近的一篇文章[Lawrence 98]指出，接近 80%的 web 已经是隐式 web。如果我们要从隐式 web 获取数据，则必须开发技术来识别产生 web 页面的站点，对这些站点进行分类，自动对它们创建查询接口。

将来的工作除了继续提高 web 信息提取的性能、使之具有更广阔的适应性外，更多的工作要放在 web 信息集成整个系统上来。

致谢

本文的工作是在复旦大学吴立德教授和黄萱菁副教授的指导下进行的，也得到了富士通研究开发中心的石崎洋之、瞿有利、胡国昕的支持与帮助，在此一并表示衷心的感谢。

参 考 文 献

[Ashish 97] Naveen Ashish and Craig A. Knoblock. Wrapper generation for semi-structured internet sources. *SIGMOD Record*, 26(4):8-15, 1997.

[Chawathe 94] S. Chawathe, H.Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, J. Widom The TSIMMIS Project: Integration of Heterroteneous Information Sources. In *Proceedings of IPSJ Conference*, page 7-18, Tokyo, Japan, October 1994.

[Craven 98] Mark Craven. Learning to extract symbolic knowledge from the world-wide web. In *Proc. of the AAAI Fifteen National Conf. on Artificial Intelligence*, 1998.

[Kushmerick 97] N. Kushmerick, R. Doorenbos, and D. Weld. Wrapper induction for information extraction. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, 1997

[Lawrence 98] S. Lawrence. Searching the world wide web. *Science*, 280(4):98-100, 1998.