

# 基于 B2B 垂直搜索的网页信息抽取系统研究

刘 丹<sup>1</sup>, 崔 阳<sup>2</sup>

(1. 南海舰队司令部, 广东 湛江 524001;  
2. 中国劳动关系学院, 北京 100048)

**摘 要:**为了解决从网页中准确抽取产品信息这一 B2B 垂直搜索引擎的关键问题,以站点树为模型,首先分析了企业网站的结构特征,在此基础上构建了一个面向 B2B 垂直搜索引擎的网页信息抽取系统。该系统利用站点树在企业站点大量网页中识别出产品页,并进行去噪处理,然后使用基于规则的方法抽取产品页中包含的产品描述信息和参数信息。通过该系统抽取到的各类产品信息较为准确,且效率得到明显提高,适用于 B2B 垂直搜索引擎中对产品的描述、分类及搜索。

**关键词:**B2B 垂直搜索;网页信息抽取;企业站点树;去噪

**中图分类号:**TP393.09

**文献标识码:**A

**文章编号:**1673-629X(2013)02-0153-04

**doi:**10.3969/j.jssn.1673-2013.02.041

## Research on System of Web Information Extraction Based on B2B Vertical Search Engine

LIU Dan<sup>1</sup>, CUI Yang<sup>2</sup>

(1. Headquarters of the South China Sea Fleet, Zhanjiang 524001, China;  
2. China Institute of Industrial Relations, Beijing 100048, China)

**Abstract:**To solve the problem of information extraction on web pages, which is one of the key technologies of B2B vertical search engine, taking website as model, structure of the corporation website is analyzed firstly, based on which a system of web information extraction for B2B vertical search engine is constructed. The website tree is used in the system for identification and noise elimination of the product pages, and then description and parameter information of the products contained in product pages are extracted according to the rules. All kinds of information extracted accurately and efficiently by the system can be used for description, classification and searching of the products in B2B vertical search engine.

**Key words:**B2B vertical search engine; web information extraction; corporation website tree; noise elimination

## 0 引 言

垂直搜索是针对某一行业、某一领域或某一主题而进行的专业搜索,是综合搜索技术的深化。B2B (Business to Business)指企业间通过互联网进行产品、服务及信息交换,是电子商务的重要组成部分。企业用户在使用 B2B 平台进行一次特定的商业交易时,通常只关注某个类别、某个品牌的一项商品的详细信息,这就要求 B2B 领域使用的搜索引擎必须能够快速、直接、准确地查询并呈现用户当前最需要的信息。因此 B2B 成为垂直搜索的一个重要应用领域。

B2B 的发展与垂直搜索技术的进一步完善密不可分,而网页信息抽取又是垂直搜索的关键问题之一。因此文中重点研究适用于 B2B 垂直搜索的网页信息

抽取系统。

## 1 B2B 垂直搜索中的网页信息抽取

垂直搜索又称主题搜索,即用户对信息的需求往往针对受限领域和面向特定的主题,搜索的结果具有对信息的分类细致精确、数据全面深入、更新及时、重复率低等特点。垂直搜索引擎的兴起与发展,实际上代表了当前 WEB 搜索技术正在从以 Google 和 Baidu 为代表的“通用化”、“扁平化”的信息搜索方式向以“个性化”、“智能化”和“垂直化”为特征的搜索方式发展<sup>[1]</sup>。

垂直搜索引擎在信息的采集、加工和处理方面都有特殊之处。首先,垂直搜索引擎的信息采集以深度优先为策略,这是因为垂直搜索引擎以解决某一领域或专业问题的搜索为目的,为了避免遗漏有价值的信息,必然要求信息采集具有足够的深度<sup>[2]</sup>。其次,

收稿日期:2012-05-28;修回日期:2012-08-30

基金项目:中央高校基本科研业务费专项基金项目(12zy019)

作者简介:刘 丹(1976-),男,硕士,研究方向为数据挖掘。

垂直搜索引擎对采集的信息要进行加工,包括结构转化、去噪、分类等。第三,在信息检索方面,垂直搜索引擎不仅能够对网页信息中的结构化信息进行检索,还能提供结构化与非结构化信息相结合的检索方式,且搜索结果更加多样化,如按时间排序、按相关度排序、按某个结构化字段排序等等<sup>[3]</sup>。

可以将垂直搜索引擎大体分为三个功能部分(见图1):网页下载、网页信息抽取和网页检索。网页下载是垂直搜索引擎构建的第一步,负责从互联网或特定网站的网页中识别、下载和保存与搜索主题相关的网页。网页信息抽取是垂直搜索引擎的核心部分,是将非结构化或半结构化的网页数据转换为特定的结构化数据,并从中抽取与搜索主题相关的信息。网页检索则是设计查询界面和查询算法,根据用户提交的查询请求在索引文件中进行查找,并将查询到的结果呈现给用户<sup>[4]</sup>。

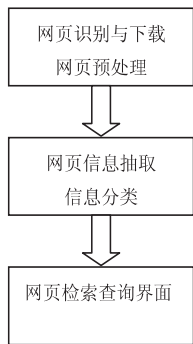


图1 垂直搜索引擎基本功能模块

就 B2B 垂直搜索引擎而言,其网页信息抽取系统必须要满足准确、高效抽取产品相关信息的要求。如果 B2B 垂直搜索引擎在含有大量噪音的企业产品信息中进行查询,则有可能造成查询速度下降、查准率降低,甚至导致用户遗漏重要信息。此外,抽取到的产品信息还将作为产品分类的基础数据集。如果分类结果不够准确,可能会使用户在搜索某一类产品信息时,出现与该类产品毫不相关的其它产品信息。例如在搜寻品牌名为“APPLE”的电子产品时,搜索结果中混有“苹果”这种水果的产品信息。

构建网页信息抽取系统的主要过程是:首先从特定企业网站中识别出包含产品信息的网页(简称为产品页),并将其下载、去噪和存储;然后从经去噪和转换的网页数据中抽取出产品的相关信息。基于此,可将构建过程分为产品页识别与去噪,以及产品页信息抽取两阶段。

## 2 产品页的识别与去噪

数据量是 B2B 垂直搜索引擎的基础之一。如果数据量不能达到一定量级,则 B2B 搜索引擎提供的搜

索结果难以全面和实用<sup>[5]</sup>。目前,B2B 垂直搜索引擎的数据主要有通过引擎下载和由企业提供两种类型,其区别在于前者由垂直搜索引擎主动寻找并下载与搜索主题相关的网页,后者是企业根据自身实际需要直接将数据发布到 B2B 垂直搜索引擎平台。B2B 垂直搜索引擎的网页信息抽取系统主要针对的是第一种类型。

### 2.1 企业站点层次树的建立

当前大多数生产企业都建有网站,用于企业自身宣传和发布、推广企业产品。企业站点中包含的信息是多方面的。一般来说,当从互联网中找到一个特定的企业站点时,B2B 垂直搜索引擎关注的是站点中以含有该企业产品名称、外观、功能、价格等信息为主的产品页,而对诸如招聘信息、联系方式等网页较少或完全不关注。也就是说,B2B 垂直搜索引擎的网页信息抽取系统需要的仅是企业站点中所有网页的一部分。

怎样识别企业站点中哪些网页为产品页,是一个较为困难的问题。有一种方法是对站点中所有网页的布局进行比较,将大量布局相同或相似的网页视为产品网页。这种方法的准确性较高,缺点是过于依赖站点自身的网页状况,如果某一企业站点各个产品页布局较为混乱,则有可能无法正确识别产品页。另外,这种方法有可能将与产品页布局类似的无关网页,如产品的列表页,也识别为产品页。另一种方法是通过人工确定一些识别规则,如认为网页的 URL 中含有“PRODUCT”关键词的即为产品页。但这种方法的关键词需要完全由人工给定,而且由于关键词的易变,使方法难以保证产品页识别的准确性和完整性。

文中采用企业站点树模型<sup>[6]</sup>,通过网页间的链接关系来识别产品页。方法是首先通过遍历企业站点的网状链接结构,生成一棵如图2所示的企业站点树,树节点表示站点中各网页的链接;然后利用站点树来识别产品页。在建树过程中,以企业站点首页作为树的根节点,首页中的各链接作为其子节点,以此类推。

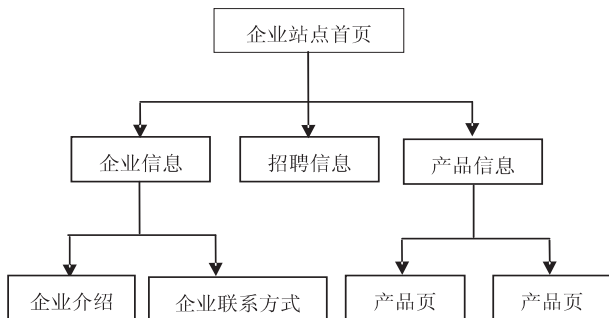


图2 企业站点树结构

为保证生成的站点树为有向无环树,并将大量重复链接去掉,规定某一链接仅在首次出现时被记录。此外,由于一般情况下产品页都在企业站点第三层或

第四层出现,因此企业站点树深度设置为3或4,即当站点树的第三层或第四层节点生成后,对企业站点的遍历即可停止。

从图1可以看出,企业站点树的首页地址即为树根节点。根节点之下的第二层节点通常是由首页导航栏引出的各主题,如企业信息、产品信息等等。有些主题可能没有下层节点,但对于产品信息这样的主题,其下层节点可能直接表示各个产品页;也有可能表示产品的列表页,而产品页是这些列表页的下层节点。经过分析大量企业站点发现,如果树中某一节点的各子节点表示产品页,则该节点的出度会远大于其它节点。因此在建立企业站点树时,要计算并保存每一节点的出度,将出度超过阈值的节点之下所有子节点表示的网页下载。当阈值设定较合理时,这些下载的网页为产品页的准确率超过90%,仅有少量噪音网页在允许的范围内存在。经过实验测定表明,阈值为8时达到较好的效果。这样就不需要将企业站点的全部网页都下载,从而提高了B2B垂直搜索引擎的效率。

### 2.2 产品页噪音数据的清洗

一个产品页中除了产品的各项信息外,通常还包括维系网页间关系的超链接、提示信息、广告等内容。从抽取产品的角度看,这些都可视为噪音数据。此外,产品页中的噪音数据还应包括用标记语言或脚本语言定义的网页样式,如CSS等,统称为低端噪音数据。

对于网页低端噪音数据的清洗较为简单,主要基于网页源代码进行。在网页源代码中找出标记语言或脚本语言的相应标签,可以很容易地将这些噪音数据过滤,得到“清洁”的产品页。

## 3 产品信息抽取

在完成产品页的识别和去噪后,即可开始产品信息的抽取。B2B垂直搜索引擎需要抽取的信息可大体分为三类:产品名称、产品描述和产品参数。其中产品名称可在产品页下载时根据产品页链接的锚文本获取。产品描述信息应包括两部分:图片和描述文本。图片的抽取在产品页下载时已经完成,只须将图片与锚文本的对应关系记录即可。因此产品描述信息和产品参数信息是B2B垂直搜索引擎的网页信息抽取系统抽取的主要对象。这主要是因为:首先,B2B垂直搜索引擎需要使用这些信息作为搜索的关键词;其次,用户在得到查询结果列表后,一般还要进一步掌握感兴趣的产品的各项信息,如产品价格、颜色、体积等,以及产品功能、性能、其它用户评价等;第三,抽取到的信息还将用于分类、聚类等,使搜索引擎能够实现对新产品的自动分类<sup>[7]</sup>。

由于网页使用的HTML标记不包含任何语义,因此以HTML语言所表述的Web网页不适合作为一种数据交换方式由机器处理<sup>[8]</sup>。目前多数网页信息抽取方法仍是以文档对象模型DOM为基础的,主要包括以下几种方法:

(1)基于规则的信息抽取。这种方法主要通过观察大量网页,找出要抽取的信息中包含的共同规律,然后经人工制定或通过标注过的样本学习来生成规则,如规则分级制思想<sup>[9]</sup>等。这种方法目前使用最为广泛。

(2)基于自然语言的信息抽取。这种方法较为有效,但要求大量实例训练,且处理速度比较慢。典型的如基于语义分析的网络蜘蛛方法<sup>[10]</sup>等。

(3)基于本体的信息抽取方法。这种方法根据领域特点设计出本体的数据框架,抽取方法与处理的网页格式无关,应用效果较好,但其设计较困难。如基于本体的表格信息抽取方法<sup>[11]</sup>。

怎样从半结构化、非结构化的Web网页中抽取信息,是当前Web数据源集成所遇到的重大挑战,也是B2B垂直搜索引擎构建面临的一个难题。文中采用基于规则的信息抽取方法。这种方法主要通过观察大量网页,找出要抽取的信息中包含的共同规律,然后经人工制定或通过标注过的样本学习来生成规则,以指导信息抽取。该方法的效率较高,但存在需要人工干预,且抽取过程和结果高度依赖预先制定的规则等缺点。由于企业站点的产品页特征较为近似,因此只须选取适当的规则表示方法,仍可以获得较好的效果。

### 3.1 产品描述信息的抽取

一个典型的产品页布局大体如图3所示。

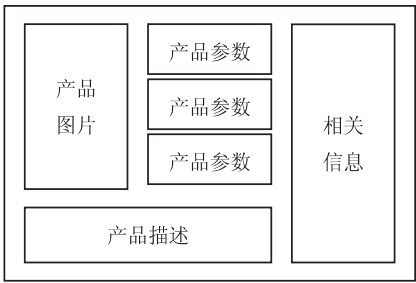


图3 产品页布局

一般情况下,产品页中的各种文本信息都包含在网页源文件的<DIV>或<TABLE>标签中,称为TEXT。文中采用了文本块权值计算的方法来抽取产品描述信息。首先遍历产品页DOM,抽取其中所有的文本块,然后计算每个文本块的权值。在遍历产品页的DOM结构时,如果<DIV>和<TABLE>中有嵌套,则提取最内层的TEXT值。将所有文本块中权值最大的称为最大文本块,这个最大文本块即被认为是产品描述信息文本块。

首先使用一个产品描述信息的训练集将描述信息中出现频率超过预设阈值的词找出并保存在一个词权值表中,并根据词的出现频率,按一定的方法分别赋予一个权值。在计算文本块权值时,如果文本块中含有词权值表中的词,则为文本块加上该词所对应的权值。除正文词外,根据产品描述信息中标点符号、换行字符相对较多的特点,统计文本块中出现的标点符号和换行字符总数,再乘以一个权值。此外文本块的长度也是一个需要考虑的重要因素,同样要赋予权值。这样,文本块权值就为三部分权值之和。

设  $W_p$  为词权值;文本块长度为  $TL$ ,对应的权值为  $W_{TL}$ ;标点符号和换行等字符总数为  $SC$ ,对应的权值为  $W_{SC}$ 。则文本块权值  $W_T$  为:

$$W_T = \sum W_p + TL \times W_{TL} + SC \times W_{SC}$$

词权值表可以看作是一种指导产品描述信息抽取的规则。各项权值可根据实验值和抽取要求设定。

### 3.2 产品参数信息的抽取

产品参数信息是用户直观了解产品的重要数据。如 MP3 应包括容量、颜色、价格、重量、屏幕尺寸、分辨率等多种参数。文中采用框架表示法<sup>[12]</sup>来表示指导产品信息抽取的规则。通常,产品各项参数以“属性-值”的形式出现,可称之为键值对。仍以 MP3 为例,所需的信息用框架表示的规则形式如下:

框架:<MP3>

品牌:范围(爱国者,三星,苹果)

缺省:爱国者

容量:单位(GB)

颜色:范围(银,红,蓝,白)

缺省:银

重量:单位(克)

价格:单位(元)

由于某一企业站点中的产品页格式在一定时期内是稳定的,因此仅通过分析少量产品页即可给出该站点产品页参数信息的框架。在分析一个产品页的 DOM 时,就可以将框架中对应的各项参数信息抽取并保存。如果产品页格式发生变化,则仅需将变化的部分在框架中修改即可。

框架表示法的缺点是在抽取不同企业站点的产品页参数信息时需要不同的框架。但由于各企业站点产品页的布局具有共性,因此框架的生成较为简单,系统运行结果显示不会对产品参数信息抽取造成明显影响。

## 4 实验

首先从互联网中随机选取了 20 家化工企业的企

业站点,根据企业站点树遍历方法,将每一站点的产品页识别并下载。

以 <http://www.xingyue-zskt.com> 为例,其生成的企业站点树结果如图 4 所示。从图中可以看到,站点中大量无关链接已被过滤,第二层节点 <http://www.xingyue-zskt.com/products.php> 下包含了较多子节点,而这些节点正是表示产品页的。

```
Input the URL:
http://www.xingyue-zskt.com

The tree of the website is:
level 1 http://www.xingyue-zskt.com
level 2 http://www.xingyue-zskt.com/about.php
level 2 http://www.xingyue-zskt.com/products.php
level 3 http://www.xingyue-zskt.com/proDetail.php?id=1
level 3 http://www.xingyue-zskt.com/proDetail.php?id=3
level 3 http://www.xingyue-zskt.com/proDetail.php?id=25
level 3 http://www.xingyue-zskt.com/proDetail.php?id=187
level 3 http://www.xingyue-zskt.com/proDetail.php?id=188
level 3 http://www.xingyue-zskt.com/proDetail.php?id=189
level 2 http://www.xingyue-zskt.com/service.php
level 2 http://www.xingyue-zskt.com/feedblack.php
level 2 http://www.xingyue-zskt.com/contact.php
level 2 http://www.xingyue-zskt.com/enabout.php
level 3 mailto:george.wu@xingyue-zskt.com
level 2 http://zsdongying.cn.alibaba.com/athena/bizreflist/zsdongying.html
level 2 http://china.alibaba.com
level 2 http://zsdongying.cn.alibaba.com
```

图 4 企业站点树的运行结果

将 <http://www.xingyue-zskt.com/products.php> 下的各产品页下载,去掉低端噪音数据后保存。选取其中一部分产品页,将其中包含的产品描述信息作为训练集,用于建立词权值表。例如,初始的词权值表中可能首先包含“说明”、“性能”、“功能”等词。按其在训练集中出现的频率,赋予相应的权值。然后使用这一词权值表计算产品页各文本块权值并提取最大文本块作为产品的描述信息。如果在找到的最大文本块中发现了新的出现频率较高的词,则将其存入词权值表,进一步指导产品描述信息的抽取。

同样,通过观察选取的产品页页面特征,可以找出产品参数信息的各键值对,如“价格-元”、“重量-克”等,从而生成框架。然后利用这一框架来指导产品参数信息的抽取。实验表明了这种信息抽取方法取得的效果比较理想。目前该信息抽取系统已经在北京九城网络有限公司的 TOOTOO 垂直搜索引擎上得到了应用。

## 5 结束语

B2B 垂直搜索引擎的性能直接影响到企业通过 B2B 方式交易的信心和效率。文中提出了一种面向 B2B 垂直搜索引擎的网页信息抽取系统,对产品页识别、去噪和产品各项信息的抽取等一系列问题进行了研究,并得到了应用。目前,产品页的信息抽取仍以网页的 DOM 结构为基础,缺乏对信息本身关联性的分析和利用。即采集来的信息量丰富庞大,但对其缺

(下转第 161 页)

对于同样的一段性能数据采集,等时间间隔采集方法在[200,1000]上共采集数据54次,而文中的算法在此段区间共采集数据40次,可以看出,在拟和曲线精确度相同的情况下,文中的算法,可以节约宝贵的网络带宽。另外,本算法还可以由用户自己来配置参数,可以使得对不同的网络性能参数进行灵活的配置,从而使此算法具有较大的适用范围。

3 结束语

当前,网络规模的不断扩大、功能复杂性的不断增加,给网络管理带来了前所未有的挑战。目前现有的大量网络管理系统软件中,一般采用的数据采集算法都是等时间间隔的数据采集算法:由系统或者用户事先设定好采样时间间隔,周期性地对性能数据进行采集和分析。针对传统的等时间间隔的数据采集算法,在占用网络带宽和拟和精确度上存在的缺陷,文中提出的自适应数据变化规律的数据采集算法在网络性能管理中对数据进行拟和时,可以在较小的管理通信开销下,得到更加精确的拟和曲线,有利于网络的健康发展。

参考文献:

[1] 王平,赵宏,陈海涛,等. 一个基于 SNMP 的简单网络管理系统的设计与实现[J]. 小型微型计算机系统,2001,22(9):1047-1050.

[2] 李木金,王光兴. 一种被用于网络性能管理的模型及实

[J]. 计算机学报,1999,22(11):1196-1203.

[3] 薛静,樊蓉,郑玉山. 基于回归分析的网络性能管理[J]. 微电子学与计算机,2005,22(2):161-164.

[4] 倪娜,晏蒲柳,夏德麟. 基于统计分析的网络性能管理与故障预测[J]. 计算机工程,2003,29(6):64-66.

[5] 杨军,张德运,张云翼. 基于分簇的无线传感器网络数据汇聚传送协议[J]. 软件学报,2010,21(5):1127-1137.

[6] Wu Y W, Li X Y, Liu Y H, et al. Energy-efficient wake-up scheduling for data collection and aggregation[J]. IEEE Transactions on Parallel and Distributed Systems, 2010, 21(2):275-287.

[7] Lin C, Tian Y, Yao M. Green network and green evaluation: mechanism, modeling and evaluation[J]. Chinese Journal of Computer, 2011, 34(4):593-612.

[8] Hao Z, Schizas L D, Giannakis G B. Power efficient dimensionality reduction for distributed channel aware Kalman tracking using WSNs[J]. IEEE Transactions on Signal Processing, 2009, 57(8):3193-3207.

[9] Jiang H B, Jin S D, Wang C G. Parameter-based data aggregation for statistical information extraction in wireless sensor networks[J]. IEEE Transactions on Vehicular Technology, 2010, 59(8):3992-4001.

[10] 许斌. JXTA-JAVA P2P 网络编程技术[M]. 北京:清华大学出版社,2003.

[11] Oaks S, Traversat B, Gong Li. JXTA 技术手册[M]. 朱剑平译. 北京:清华大学出版社,2004.

[12] 丁晓贵,刘桂江. 基于 SOPC 的远程数据采集系统设计[J]. 计算机技术与发展,2010,20(1):229-231.

+++++ (上接第 156 页)

乏深层次的加工、处理和运用。造成这种情况的重要原因之一即是分析和处理信息的手段尚显不足。今后工作中将考虑利用信息间的关联知识和因果知识来指导产品信息抽取的方法。

参考文献:

[1] 刘畅. 综合搜索引擎与垂直搜索引擎的比较研究[J]. 情报科学,2007,25(1):97-102.

[2] Cai Deng, Yu Shipeng, Wen Jirong, et al. Extracting content structure for Web pages based on visual representation[C]// Proceeding of the 5th Asia Pacific Web Conference. Berlin: Springer-Verlag, 2003:406-417.

[3] 林文清. B2B 垂直搜索引擎在信息获取技术中的应用[J]. 情报杂志,2007(9):120-121.

[4] 余森,杨丹,赵俊芹. 垂直搜索引擎的关键技术研究[J]. 软件导刊,2007(12):31-33.

[5] 赵金仿,赵艳,缪建明. 网页信息抽取及其自动文本分类的实现[J]. 计算机技术与发展,2008,18(10):37-39.

[6] Cui Yang, Yang Bingru. A Method of Eliminating Noisy Infor-

mation in Web Pages for B2B Vertical Search Engine[C]// Proceedings of 2008 International Workshop on Information Technology and Security. [s. l.]:[s. n.], 2008:990-993.

[7] Caulkins J P, Ding W, Duncan G. A method for managing access to web pages: filtering by statistical classification (FSC) applied to text[J]. Decision Support Systems, 2006, 42:144-161.

[8] 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.

[9] 李向阳,戴江山,张亚非. 一种 Web 信息抽取规则的优化方法[J]. 兰州理工大学学报,2006,32(1):90-93.

[10] 罗立宏,陈志. 基于语义分析的垂直搜索网络蜘蛛[J]. 计算机工程与设计,2008,29(18):4662-4665.

[11] 周明建,高济,李飞. 基于本体论的 Web 信息抽取[J]. 计算机辅助设计与图形学学报,2004,16(4):535-541.

[12] 王永庆. 人工智能原理与方法[M]. 西安:西安交通大学出版社,2001.

# 基于B2B垂直搜索的网页信息抽取系统研究

作者: [刘丹](#), [崔阳](#)  
作者单位: [刘丹\(南海舰队司令部, 广东 湛江 524001\)](#), [崔阳\(中国劳动关系学院, 北京 100048\)](#)  
刊名: [计算机技术与发展](#)  
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2013 (2)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201302041.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201302041.aspx)