基于多维语义的互联网药品信息提取方法①

顾轶灵

(复旦大学 软件学院，上海 201203)

摘 要：提出了基于多维语义的互联网药品信息提取方法，构建语义词典通过从多个维度对互联网药品知识进行描述，克服了不同来源网页之间的异构性并找出了其隐藏的共性。同时，采用了基于结构语义熵的方法对目标网页信息聚集区域进行定位，从中提取感兴趣的药品信息。最后再通过语义词典对提取的信息进行验证并自动生成 XPath 提取规则进行补充。该方法能够自动有效地从互联网的多个信息来源获取药品信息，实验证明其具有较高的准确性与召回率，可以为政府相关部门加强互联网药品市场监管提供足够的信息依据。

关键词：Web 信息提取；多维语义词典；互联网药品信息；结构语义熵；XPath

Multidimensional-Semantics-Based Web Medicine Information Extraction

GU Yi-Ling

(Software School, Fudan University, Shanghai 201203, China)

Abstract: A multidimensional-semantics based Web information extraction method is proposed in this article to extract medicine information on the Web. The method overcomes the heterogeneity of Web pages from different sources and finds the common characteristics among them by building up a semantic dictionary and describes the knowledge of medicine information over the Web. At the same time, it utilizes a structural-semantic-entropy-based approach to detect data-rich sections on Web pages, then extract information of interest from them and finally verify and supplement the extracted information by generating extraction rules using XPath. The method is able to obtain information from heterogeneous sources both automatically and effectively. Experiments shown that it has high precision and recall, thus can provide sufficient information for the government to enhance supervision of medicine market on the Web.

Key words: Web information extraction; multidimensional semantic dictionary; Web medicine information; Structural-semantic entropy; XPath

1 引言

随着近年来电子商务产业的迅速发展，互联网药品市场也在日益增长。然而，随之而来的隐患也在不断加剧：由于许多未经国家药监局批准的网站擅自在线销售药品，更有甚者利用互联网的隐蔽性进行假药的销售，就连具备互联网药品信息发布或交易资质的网站发布的信息也常常不够准确规范，导致整个互联网药品信息市场鱼龙混杂，普通消费者很难辨别药品信息的真伪，用药安全受到严重的危害。

对于这样严峻的现状，政府相关部门急需加强对互联网药品市场的监管力度，但是由于互联网上的信息量庞大，传统的人工监测手段无法跟上互联网药品信息的快速增长，准确自动的智能监测手段就成为了非常迫切的需求。如何由机器自动获取网页并且识别出其中可能包含的药品信息无疑是自动监测之中最首要的一个问题，而事实上这正是一个 Web 信息提取的典型问题。

Web 信息提取技术目前已经受到广泛关注，它的目标是将 Web 页中的非结构化或半结构化的数据信息转换为结构化的信息重新存储，便于进行查询和利用。但在实际应用中这种技术尚很不成熟，手工构建基于网页标签表达式的正则模板是最常用的提取手段。但

① 收稿时间:2011-03-10;收到修改稿时间:2011-04-18

是这种方法有很大的局限性，无法自动适应各个目标网站在页面结构上的更改，而且也无法识别越来越多的新增网站。

为了能够准确全面地自动识别互联网上的药品信息，本文提出了一种基于多维语义的互联网药品信息提取方法，从多种角度对互联网药品信息的基本形式进行描述，而不依赖信息在网页上的具体表现形式，从而达到在不同来源的网页中自动识别药品信息的目的。从实验结果来看，该方法具有较高的准确率与召回率，可以在实际应用中起到较好的效果。

2 相关工作

Web 信息提取是在网页(主要是 HTML 页)中的半结构化数据里，找到用户感兴趣的内容，将其转换成结构化的数据，方便进行查询、数据统计与挖掘的过程。而网页具体的表现形式又有着很大的异构性，为了克服这些异构性，为每种结构生成一个通用的提取模板(被称为 Wrapper，包装器)人们作了大量的研究，目前的主要方法可以分为下面几个类别：

1) 人工构建包装器的提取方法。这种方法要求用户根据目标网页结构利用一些特定语法或是脚本语言(如 Perl)来人工编写包装器规则，其用户往往需要一些编程知识，提高了该方法的使用门槛。比较典型的此类 Web 信息提取系统有 TSIMMIS[1]、Web-OQL[2]以及 WRAP 等[3]。

2) 人工监督的包装器归纳(Wrapper Induction)。这种方法要求用户提供一个预先标注好的样本集，系统经过对标注样本的自动归纳，找出其中的规律后输出特定的包装器。这样，普通用户可以在系统提供的图形界面中对样本进行标注，与前一类方法相比节约了成本。这类 Web 信息提取系统比较具有代表性的有 WHISK[4]、WIEN[5]和 STALKER[6]等。

3) 半监督的信息提取方法。这类系统往往通过对网页结构进行分析，寻找重复模式，提取中其中发现的数据。与人工监督的方法不同，它不需要人工标注好的样本集，而是需要用户在系统粗略挑选出数据后，人工选择其中真正需要的部分。利用这种方法的系统主要有 IEPAD[7]和 OLERA[8]等。

4) 无监督的信息提取方法。无监督的方法不需要用户提供已标注的样本集，也无需用户在包装器归纳过程中进行人工干预。但用户通常还是需要在产生的多个可能的包装器产中进行选择，或是为提取出来的数据指定相应的名称。典型的此类系统包括 RoadRunner[9]、DeLa[10]等。

从其他角度出发也有其他的分类方法，如基于 HTML 结构的方法、基于自然语言处理的方法、包装器归纳工具、以及基于本体的方法等等。

通过上述研究可以发现，现有的 Web 提取技术对于互联网上不断产生的各种形式药品信息，特别是在中文网页上并不能起到很好的提取效果。一些基于提取规则的包装器生成方法虽然有着很高的准确率，但需要大量的人工干预，一旦目标网站的页面稍微进行改动就可能失效，不断维护新的目标网站的提取规则或是标注样本更是一项非常繁复的工程。而基于网页结构分析的方法通常是通过发现页面中的重复模式从而获取其中变化的数据，但这样通常会得到很多用户不感兴趣的内容，也往往只能提取多记录的列表页面信息而无法得到单个物品页面的详细信息。同时，目前绝大多数基于自然语言处理的 Web 信息提取方法都是基于英文网页，而且通常用于处理大段自然文本，对于常以短语形式给出的半结构化的药品信息效果并不理想。

3 利用多维语义提取药品信息

3.1 方法概述

本文的方法通过设计互联网药品信息领域的多维语义词典来描述领域知识，并计算网页节点的结构语义熵来定位信息聚集区域，通过领域知识中的各种规则对信息进行提取、验证、补充，最终将得到的结构化的信息存入药品信息数据库。输入待提取的药品详细信息页面(如图 1 所示)，经过分析后可以排除页面内的干扰信息，获得属性聚集区域内描述该药品的属性名值对。

图 1 典型的药品详细信息页面

本文描述的互联网药品信息提取方法主要分为构

建多维语义词典以及基于语义词典的信息提取两部分，其中信息提取的流程主要有以下几个步骤：

1) 网页预处理；

2) 定位属性聚集区域；

3) 提取属性名值对；

4) 根据语义词典的规则进行验证与补充。

3.2 多维语义词典的构建

互联网药品信息多维语义词典为整个信息提取方法提供领域知识，是进行信息提取的依据。由于不同网站的互联网药品信息从语言表述到网页结构等方面都存在很大的差异，所以如何描述出这个领域内的通用知识从而屏蔽这些差异给信息提取带来的困难成为了整个方法的关键环节。本文基于对当前互联网药品交易网站的调研，在以下多个维度上对整个互联网药品领域的语义信息进行了描述：

1) 通用语义。为某些特定的概念定义了词汇集合，比如表述度量单位的词汇集 measure:{ml,mg,粒,毫升,克,盎司,支,盒,个…}、表述数字的词汇集 number:{0,1,2,3,4,5,6,7,8,9,一,二,三,四,五,六,七,八,九,十,百,千,万}以及表述省份的词汇集 province:{京,津,沪,渝,苏,浙,…}等等。

2) 药品属性语义。由于在网页中药品信息是以属性名值对的形式给出的，所以属性语义可以很好地描述这类产品，所以属性的定义就显得非常重要。属性语义分为两个主要部分：属性名称和属性值。对于属性名，由于属性名及其表达方式是有限的，给出其可能的正则表达式就可以在网页中成功匹配绝大多数属性名。而属性值却是无法穷举的，其中有一部分属性值带有一定的规律，可以使用正则表达式或词汇集组成的通用模板来表述，例如批准文号的为“国药准字[0-9a-zA-Z]+”以及“{province}卫食证字（\d{4}）第[\d-]+号”等，从而能够匹配类似“国药准字 H20010115”和“苏卫食证字（2009）第 320124-010001 号”之类的值，而“规格”的表达式则可以是“{number}{measure}([*xX]{number}{measure}”从而可以匹配“40mg*10 粒”这样的取值。另一部分属性值没有表达式可以精确描述，比如药品的商品名称如“泰诺”、“开瑞坦”等或是药品的，对于这样的属性，定义了一个实例词库，通过从国家药监局官方网站使用页面正则表达式模板提取“产品名称”以及“通用名称”等词库，当在后文提到的基于启发式方法进行属性提取失败时，可以使用词库来进行补充。

3) 属性环境语义。虽然在不同网站的具体 HTML 结构中，药品属性的周围的标签、嵌套结构都有所不同，但也能在其中找到一些共同的特点。比如，药品的名称在我们调研到的所有网站中，均处于属性名值对列表的最上方，所以可以为其指定可能出现的位置如 BeforeRegion、FirstChildOfRegion、FirstChildOfParent 等。同时，一些常用的 HTML 特征虽然不能直接匹配属性，但可以帮助提高属性识别的效率，比如 HTML 标题节点<h1>~<h6>经常包含名称信息，而<strike>或<del>元素(浏览器渲染时会添加删除线)常用来表征“市场价”。

4) 药品类语义。药品类的定义中包含属性的基数、属性间的约束条件等。例如，“白加黑”可以有两个“批准文号”，因为其白片和黑片分别是两种不同的药物，故“批准文号”就可以有多个属性值。另外，为药品定义了虚拟属性如“标题”，这样的属性虽然不是药品本身的属性，但在互联网药品销售时所有网站都会使用“商品名称”或“通用名称”或是将其组合起来作为药品标题进行表述。所以我们给这个虚拟属性定义了“{hasProductName}/{hasGeneralName}, {hasProductName }, {hasGeneralName}”这样一个顺序的表达式列表，在非虚拟属性提取完毕后，依次尝试组合出可用的属性值。当其中的某个引用的其他属性未能提取到可用的取值时，再尝试列表中的下一个表达式，直到某个表达式中的所有属性引用都可成功提取。当所有表达式都无法成功产生可用值时，可以根据环境语义中定义的位置、HTML 特征信息结合指定的取值词库尝试进行补充提取。事实上，这个表达式列表等价于互联网药品类内的约束条件“Medicine has ProductName || Medicine has GeneralName”，即药品必须至少包含“商品名称”或“通用名称”之中的一个。这样具有组合关系的约束条件仅靠单个属性基数的定义是无法实现的。

3.3 互联网药品信息提取

3.3.1 网页预处理

网页的预处理主要有两部分，完成客户端脚本执行以及将 HTML 代码解析为 DOM 树。

由于目前有部分网站出于各种目的(有一些是为了提高响应速度，也有一些是为了屏蔽爬虫对关键内容的抓取)，在用户对网页发出 HTTP(超文本传输协议)请求后，先发送一个 HTML 页，但其中部分关键内容

使用 JavaScript 进行输出，使得在不执行脚本的情况下无法得到这些内容。MIT 的 SIMILE 项目中的一个工具 Crowbar 可以用来利用 Firefox(火狐浏览器)的开源引擎 Gecko 执行 JavaScript，并给出执行完毕的网页内容。

网页的 DOM(Document Object Model，即文档对象模型)表示能够很好地描述 HTML(或 XML)网页内容的实际嵌套结构，也是 Web 信息提取过程中常用的网页表示模型。在通过 Crowbar 得到用户实际看见的网页内容后，再利用开源的 NekoHTML 库对 HTML 代码进行解析，构建符合 W3C 接口规范的 DOM 树。图 1 中属性聚集区域对应的 DOM 树子树如图 2 所示。

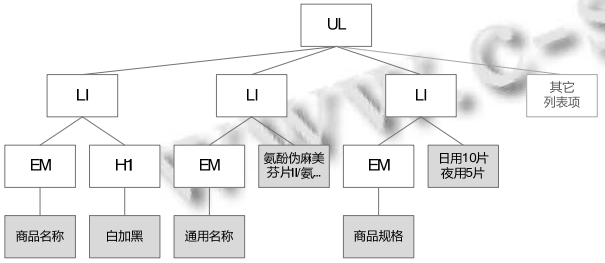

图 2 典型属性聚集区域 DOM 子树

图 2 中灰色的叶子节点均为文字节点，也就是最终显示在网页上的文字信息。这样的树形结构清晰地展现了网页内容之间的层次关系，有利于正确找出网页中的药品对象。

3.3.2 定位属性聚集区域

定位属性聚集区域等价于找出 DOM 树中具有丰富属性名值对信息的最小子树。在实际的网页中，除了当前页面所描述药品信息外，同时给出一些同类药品、促销药品的信息也很常见。这类信息一般以列表的形式给出，包含的药品信息较少，而真正的提取目标区域包含了较多的属性名值对信息。使用“结构语义熵”[11]概念，可以在同时考虑了页面结构和药品语义的情况下表征一个 DOM 树子树的属性类型的丰富程度。

首先，对 DOM 树中的每个文本节点(<img>标签中的 title 属性也可视为文本节点)，尝试匹配多维语义词典中的属性名或属性值模式。为每个匹配到的节点分配一个语义角色，如“批准文号的属性名”、“批准文号的属性值”、“生产厂商的属性名”等。

对于以节点 N 为根节点的子树，受到信息熵概念的启发，其结构语义熵可定义如下：

$$H(N) = -\sum_{i=1}^{m} p_i \log(p_i)$$

其中 pi 为某一种语义角色在 N 下的所有文本节点中出现的比例，m 为 N 下包含的语义角色的数量。未被语义匹配的节点的语义角色视为“未知”。某个结构语义熵值越大，说明在此节点下的语义角色类型越丰富。对于文本节点，其结构语义熵值为 0。对于列表，虽然可能包含很多属性节点，但其语义角色类型却不如属性聚集区域丰富，所以选取适当的结构语义熵阈值，即可将列表等干扰信息排除。最后，计算结果高于阈值的子树被选取出来，如果其中有父子关系的子树熵值相同，将保留最小子树，作为最终确定的属性聚集区域。

3.3.3 提取属性名值对

在上一步确定的属性聚集区域范围内，利用语义匹配的结果，已经可以得到一部分属性名值对的确切位置。对于仅匹配到属性名称的节点，可以采用启发式方法进行属性值提取。主要的原则基于如下的启发式假设：

1) 属性值紧跟着属性名称之后出现；

2) 某个属性的属性名称与属性值之间不会包含其他属性。

由于用户阅读网页区域的顺序(从上到下、从左到右)正对应于 DOM 树的深度优先遍历访问文本节点的顺序，所以按照此顺序即可为已知的属性名称提取相应的属性值，而由假设 2)可知如果两个属性名称之间没有非空的文本节点，那么前一个属性的属性值是缺失的。

3.3.4 根据语义词典进行验证与补充

由于不同网站选择给出的属性集合是不同的，所以药品信息中缺失部分属性是一个普遍的现象。除去网页中确实缺失的属性，有一部分属性值不能用模式精确描述，也没有匹配到相应的属性名称，但的确已在网页中已经给出(比如药品的名称)，是无法用上面的方式获取的。

在药品类语义中定义了药品属性必须包括“商品名称”和“通用名称”中的至少一个以组成虚拟属性“标题”，故若之前环节未能提取到名称信息时，将利用环境语义信息以及相应属性的取词词库，来在 DOM 树中自动识别出适用于当前网站的各个页面的属性提

取规则，来补充必要的属性值。当已经识别出某个网页的属性聚集区域后，“商品名称”和“通用名称”都无法获取时，就中断当前页面的提取过程，开始自动归纳提取规则。

同一网站内获取的页面中能够定位属性聚集区域的网页均为药品详细信息页，在这些网页中，根据环境语义中定义的 DOM 树中“标题”可能的位置如 BeforeRegion、FirstChildOfRegion 等，计算这些抽象位置在具体页面中的节点位置作为候选节点，对每个候选节点中的文本在“名称”相对应的取值词库中进行匹配。同时，将每个节点的 XPath 与匹配结果记录下来。如此循环，在分析了一定数量的网页后，找出匹配率最高的 XPath，作为用来在药品详细信息页面提取“标题”的提取规则。XPath 是 W3C XSLT 标准的主要元素，可以用来在 XML 文档中表达节点路径，通过本方法提取出的 XPath 对于同一个网站来说是通用的，因为一般来说，同一个网站的同类页面均是由同一生成规则输出的。

获得 XPath 提取规则后，将利用 XPath 提取出的缺失属性连同已经成功提取的属性集合一起，作为最终的信息提取结果输出，通过语义词典和数据库字段中的映射规则，存储到数据库中。

4 实验

4.1 评价指标

实验共进行了两部分内容，分别通过人工比对来评价属性聚集区域的识别情况以及属性名值对的提取的效果。

评价 Web 信息提取系统性能的常用指标有召回率(R, Recall)、准确率(P, Precision)以及 F 值(F-measure)。

在实验第一部分中，召回率 R 定义为：正确识别的包含属性聚集区域的页面数目/实际所有包含属性聚集区域的页面数目；准确率 P 定义为：正确识别的包含属性聚集区域的页面数目/所有被识别为包含属性聚集区域的页面数目。

第二部分中，召回率 R 定义为：正确提取的属性名值对数目/输入详细信息页面中实际的属性名值对数目；准确率 P 定义为 = 正确提取的属性名值对数目/所有提取出的属性名值对数目。

而 F 值是综合考虑上述两项的一项指标，定义为：

$F = 2 \cdot P \cdot R /(P+R)$

4.2 实验数据集

使用爬虫共从药房网、百洋健康药房、金象大药房等 10 个互联网药品交易类网站抓取了 9035 个网页，其中包含属性聚集区域的详细信息页共 4456 个。

4.3 实验结果

首先根据不同的结构语义熵阈值，得到的第一部分结果如图 3 所示。

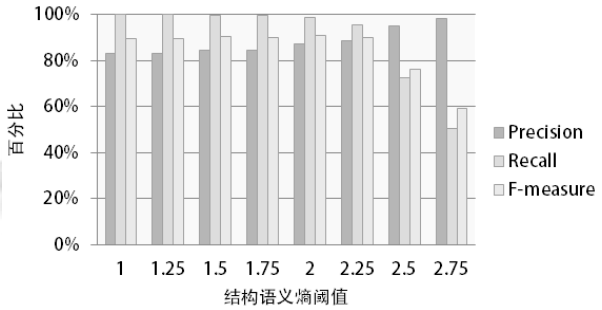

图 3　不同熵阈值下属性聚集页面识别指标

从实验结果可知，当结构语义熵阈值为 2，实验数据集的平均 F 值最高，为 95.12%。

第二部分中，对所选取的 10 个网站的共 13248 条属性的提取情况进行了人工比对，其中正确提取的有 11644 条，提取有误的占 380 条，平均召回率为 87.89%，平均准确率为 96.85%，F 值为 92.15%。

5 结论

本文通过研究互联网药品市场以及 Web 信息提取技术的现状，根据当前药品交易网站的具体情况，提出了基于多维语义的互联网药品信息提取方法。该方法通过使用从多个维度描述互联网药品领域语义的语义词典，结合基于结构语义熵的属性聚集区域定位算法，在无需后续人工干预的情况下能够达到较高的召回率与准确率，可以在互联网药品信息监测中进行实际应用。

通过分析实验中的错误结果，下一步工作将着重于多维语义词典的逐步完善以及解决生成 XPath 提取规则遇到同一网站多个模板时无法用统一的提取规则进行提取的问题。

参考文献

1 Hammer J, McHugh J, Garcia-Molina. Semistructured data: the TSIMMIS experience. Proc. of the 1st East-European

(下转第 19 页)

资产管理、网络教学、综合信息服务等系统，为学校师生提供各类信息服务。目前系统与无线网络平台的结合将成为建设的难点。创造一个开放的、支持多种应用协议并且可管理的移动学习应用平台，要求系统能全面支持高校教学、科研和办公等各种典型应用，如手机网站 web 浏览、收发邮件、文件传输、跨校区班车查询,校内地图导航、校历课程安排等，尤其是近年来迅速发展的各种多媒体应用，如实时或非实时视频点播、视频会议、即时通讯、IP 电话、教学博客和互动社区等。

5 结语

随着我校对支持移动学习无线校园网技术研究的进一步深化，通过建设支持移动学习的校园无线网络，可以给学生和老师提高自己的学习能力、研究水平创建移动学习平台，并且在原有数字化校园基础上逐渐扩充出其他新的业务应用，让每一个拥有支持移动学习终端的学生和教师都可以自如享受移动学习应用。

参考文献

1 段炳玺,李柏年,马金定.WLAN 中 IEEE802.11n 标准及关键技术研究.通信技术, 2008,8:123–125.

2 王华,李静静,何振,韩姗姗.无线 Mesh 网络技术研究.南京信息工程大学学报(自然科学版),2010,2(4):332–336.

3 陈永坚,丛林.无线网状网与应用技术.网络通讯与安全，2010,6(13):3369–3372.

4 魏晓波.无线校园建设方案及常见问题.中国信息技术教育,2009,4:86–87.

5 吴暑静.基于 IEEE802.11 无线局域网安全性研究.湖南科技学院学报,2010,31(12):86–88.

6 赵隽,黄振海,赵跃华.WAPI 与 IEEE 802.11i 安全协议通信性能分析.通信技术,2007,12(40):228–231.

7 朱智达.无线校园网建设中 QoS 部署与应用研究.广西轻工业,2009,10:85–86.

(上接第 54 页)

Symposium on Advances in Databases and Information Systems (ADBIS), 1997,1–8.

2 Arocena GO, Mendelzon AO. WebOQL: Restructuring documents, databases, and Webs. Proc. of the 14th IEEE International Conference on Data Engineering (ICDE). 1998, 24–33.

3 Liu L, Pu C, Han W. XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources. Proc. of the 16th IEEE International Conference on Data Engineering (ICDE). 2000,611–621.

4 Soderland S. Learning Information Extraction rules for Semi-structured and Free Text. Journal of Machine Learning, 1999, 34(1-3):233–272.

5 Kushmerick N, Weld D, Doorenbos R. Wrapper induction for information extraction. Proc. of the Fifteenth International Conference on Artificial Intelligence (IJCAI). 1997,729–735.

6 Muslea I, Minton S, Knoblock C. A Hierarchical Approach to Wrapper Induction. Proc. of the Third International Conference on Autonomous Agents (AA-99). 1999.

7 Chang CH, Lui SC. IEPAD: Information Extraction Based on Pattern Discovery. Proc. of the Tenth International Conference on World Wide Web (WWW). 2001,223–231.

8 Chang CH, Kuo SC. OLERA: A Semi-supervised Approach for Web Data Extraction with Visual Support. IEEE Intelligent Systems. 2004,19(6):56–64.

9 Crescenzi V, Mecca G, Merialdo P. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. Proc. of the 26th International Conference on Very Large Database Systems (VLDB). 2001,109–118.

10 Wang J, Lochovsky FH. Data Extraction and Label Assignment for Web Databases. Proc. of the Twelfth International Conference on World Wide Web (WWW). 2003,187–196.

11 吴晓彦.基于结构语义熵的互联网商品信息抽取技术研究[硕士学位论文].上海:复旦大学,2009.