基于语义的互联网药品信息抽取算法①

沈元一，郑骁庆，顾轶灵

(复旦大学 软件学院，上海 201203)

摘 要： 针对现有互联网信息抽取技术存在准确率不高、覆盖率低、人工干预多等诸多缺陷，提出了一种新的互联网药品信息抽取算法，通过引入语义技术构建三维语义词典，屏蔽不同药品信息网页在内容和结构上的异构性，同时利用所需抽取的目标药品属性信息具有一定聚集度的特征，基于信息熵的理论设计出对目标信息智能定位和抽取的方法。实验证明该算法既能降低人工干预，又具备较高的准确率和召回率。应用该算法能实时自动全面准确地获取互联网药品信息，为政府药监部门提供丰富的监管依据，对规范医药电子商务市场，保证人们的用药安全具有重要的现实意义。

关键词： Web 信息抽取；语义词典；DOM；信息熵；Xpath；医药电子商务

Web Medicine Information Extraction Algorithm Based on Semantics

SHEN Yuan-Yi, ZHENG Xiao-Qing, GU Yi-Ling

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

Abstract: This article addresses defects of current Web information extraction technology such as low accuracy, low coverage, and manual intervention required, proposes a novel extraction algorithm of web medicine information. The algorithm sets up a three-dimentional semantic dictionary by introduction of the semantics technology, masks the isomerisms of the web page contents and structures, and at the same time, taking advantage of the fact that the attributes of the target medicine tend to have a character of aggregation, designs a way of intellectually locating and extracting the target information based on the theory of information entropy. Through related experiments proves that the algorithm is able to reduce the requirement of manual intervention of the information extraction, and has a high accuracy and recall rate. The application of this algorithm can automatically, comprehensively, and accurately obtain Internet medicine information in real time, offers abundant basis of supervision for the medicine supervision department, and therefore has a significant practical meaning of normalizing medical e-business and ensuring secure medication.

Keywords: Web information extraction; semantic dictionary; DOM; information entropy; XPath; medical E-business

1 引言

近年来，国内医药电子商务发展迅速，众多生产商和药房纷纷通过互联网完成从药品信息发布更新到在线药品交易支付的整个过程，既降低了经营成本又简化了交易流程，很大程度上减轻了公众在医疗上的开支。然而由于互联网的虚拟性、隐蔽性和自组织性等特征，网上非法药店越来越多，造成目前互联网上虚假药品信息泛滥，严重危害到人们的用药安全。为了提高政府机构对医药电子商务行业的监管力度，急需对互联网上的药品信息进行全面准确实时自动的抽取和分析，给政府药监部门提供丰富的监管依据。

Web 信息抽取技术作为其中的关键技术已经有着广泛的研究基础。比较典型的 Web 信息抽取应用包括商品比价网站、舆情监测等，其大都采取为不同目标源网页手工构造相应的正则表达式匹配抽取模板的方式，尽管这样能保证较高的抽取准确率，

① 基金项目:国家科技支撑项目(2006BAH02A05-06);国家自然科学基金(60903078, 60973025)
收稿时间:2010-04-24;收到修改稿时间:2010-05-23

但人工干预的局限性也从根本上导致了其在召回率上的严重缺失，同时无法对已有网页设计和结构上的改变做出自动响应，更无法自动识别新的目标源。

为了达到全面、准确、实时、自动地抽取，我们在现有抽取技术的基础上，设计并实现了一种基于语义的互联网药品信息抽取算法，该算法不依赖于目标源网页的表现形式，而是借由语义技术专注于目标信息的内容，对源网页中的药品属性信息进行智能定位和自动抽取，既能大大降低人工干预又保持了较高的抽取准确率。同时，该算法不仅能对当前互联网上存在的药品信息进行抽取分析，还能实时地通过元搜索技术获取新的医药电子商务网站作为目标信息源，从而显著提高抽取召回率。

2 相关工作研究

HTML 网页作为互联网信息最主要的展示方式，其包含的数据是半结构化的，而文档的结构化程度越高，越有规律，计算机就越容易从同类文档中识别出相应的模式，进而更有效地理解和处理文档内容。由于 HTML 网页的内容和表现形式紧密耦合而缺乏良好的组织结构，给计算机理解和分析网页信息造成了很大困难。Web 信息抽取正是解决上述问题的技术手段，其根据网页特征生成特定的抽取规则和程序，进而将半结构化的网页文档转化为结构化的数据信息。目前 Web 信息抽取技术主要包含了如下几类：

(1) 基于包装器归纳的抽取。对网页中需要抽取的内容进行手工标注，而后由系统学习和抽取规则归纳形成包装器，再利用该包装器来抽取其它同类网页的内容。其中具有代表性的包装器归纳系统有 WIEN[1]、STALKER[2]等。

(2) 基于网页结构的抽取。利用同一个网站中数据记录的表现形式往往具有相似结构的特征，通过对这些结构进行重复模式的挖掘，找到承载这些数据的模板。典型的采取该方式抽取的系统有 RoadRunner[3]、IE Pad[4]等。

(3) 基于自然语言处理的抽取。使用自然语言处理技术对网页中的文本进行语义分析，将分析好的文本与定义好的语言模式进行匹配，以抽取其中的数据。规则模式既可以人工定义，也可以根据标注好的语料库自动学习而得。典型的采取这种抽取方式的系统有 WHISK[5]、SRV[6]等。

以准确率、召回率和人工干预程度这三项指标综合衡量，现有的 Web 信息抽取技术存在着明显的不足。基于包装器归纳的抽取方法准确率较高，但手工标注的局限性造成了其召回率的丢失，同时还需为每一个网站维护标注页面的训练集合，人工成本高且灵活性不佳，一旦数据格式被更改就会导致包装器的失效；基于网页结构的抽取方式具有较高的召回率，但由于完全基于表现形式的相似性，系统并不了解用户感兴趣的内容，因此抽取结果中可能会包含很多用户不需要的数据；基于自然语言处理的抽取方式因其完全忽略网页结构特征而专注于网页内容的语义分析，可以做到对任意网页进行处理，虽然在理论上能保证较高的召回率和准确率，但需要庞大且完备的知识库的支撑，由于人类在自然语言处理方面的研究目前尚处在起步阶段，同时知识信息爆炸又给知识库的更新带来了极大困难，所以纯粹基于自然语言处理来抽取网页信息的可行性目前还很低。

通过上述研究我们发现，现有的 Web 信息抽取技术并不能对互联网上的药品信息进行全面准确实时自动地抽取。为了尽可能减少人工干预，我们决定摒弃常用的基于包装器归纳和网页结构模式挖掘的抽取方式，引入语义技术作为突破口，通过构建互联网药品信息领域的语义词典，同时结合信息熵理论，智能识别出医药电子商务网站并对网页中存在的目标药品信息进行自动定位和抽取。虽然该抽取算法仍属于自然语言处理的 Web 信息抽取范畴，但由于我们的研究主要针对中国医药电子商务这一领域，该领域互联网药品信息发布具备一定的规范，同时药品本身又是标准化程度相对较高的特殊商品，有利于构建相对完备的领域知识库，从而保证了我们抽取算法的可行性。

3 抽取算法关键技术和设计思路

3.1 互联网药品信息三维语义词典

互联网药品信息三维语义词典是本文研究设计的基于语义的互联网药品信息抽取算法的重要基础。为了支持准确自动的信息抽取，需要对全网范围内发布的药品信息进行调研、统计和分析，充分利用其中通用的信息特征，同时尽量屏蔽不一致的信息，来保证

我们抽取算法的普适性。

通过对多家医药电子商务网站的调研分析，我们发现其在药品信息发布上有着明显的规律，即都出现了众多药品领域的常用术语和描述商品的词汇，例如“药品名称”、“生产厂商”、“批准文号”、“价格”、“成分”、“适应症”等等。这些都可以归作互联网药品信息的通用特征。然而由于互联网信息的海量，从概率统计的角度来说几乎不存在完全一致的信息发布格式。药品领域尽管标准化程度很高，但就目前而言政府机构还无法制定出相关的法律法规来充分约束和规范医药电子商务网站发布药品信息的格式，所以当前互联网上的药品信息存在不一致是不可规避的。

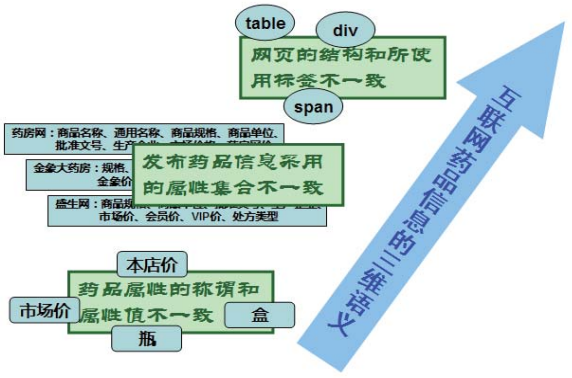


图 1 互联网药品信息的不一致

图 1 主要展示了我们所发现的三种信息不一致的情况：

(1) 药品属性的称谓和属性值不一致

主要体现在不同医药电子商务网站在表述药品属性时采用了不同的词汇或表达方式，例如一些网站描述药品名称属性使用的是“药品名称”，一些则使用“商品名称”；同一个网站描述药品价格属性时也可能存在“市场价格”、“本店价格”、“会员价格”等多种称谓。另外药品属性的取值也呈现明显的不一致，比如描述药品计量单位属性会出现“瓶”、“盒”、“支”等多种情况。

(2) 发布药品信息采用的属性集合不一致

主要体现在不同医药电子商务网站在销售药品时罗列的属性不完全相同，数量上有多有少，分类上也有差异。例如某网站选择“商品名称”、“批准文号”、“生产企业”、“配送药房”等属性，另一些网站并没有“配送药房”这一类。

(3) 网页的结构和所使用标签不一致

不同医药电子商务网站无论在网页结构还是展现风格上都千差万别，我们需要抽取的药品信息往往隐藏在众多无序的 HTML 标签之中，例如有的网站使用<table>来表现列表，有的则用<ul>、<li>替换；有的网站“药品名称”是粗体显示，有的网站“批准文号”四个字之间用空格分开。

综上所述，这三类互联网药品信息的不一致，为我们的抽取算法准确自动抽取目标信息造成了诸多困难。为了规避这些不一致，我们提出了互联网药品信息的三维语义概念，具体如表 1 所示：

表 1 互联网药品信息的三维语义概念

名称	考虑的语义	可解决的不一致
文字维	文字本身的语义，如同义词、词之间的层级关系、语法、语用等	药品属性的称谓和属性值不一致
领域维	特定领域不同词汇之间的语义关联，如多个药品属性信息组合在一起就具备了表征互联网药品信息的能力	发布药品信息采用的属性集合不一致
表现维	网页采用的结构和标签形式根据所展现信息的相似性也被赋予了一定的语义	网页的结构和所使用标签不一致

目前一些主流的通用搜索和垂直搜索引擎已开始加入对语义的支持。以百度通用搜索为例，最常用的语义技术支持服务是对搜索关键字进行语义分析，给出意思相近的词语供用户参考，其局限于文字维的范畴。部分垂直搜索比价网站，在搜索结果中也会包含一些同类商品的推荐，一定程度上利用了领域维的语义关联性。尽管语义的思想得到了一些体现，但其本质并没有系统地使用语义知识来改变其核心的搜索和抽取算法，没有从多个维度进行语义判别和约束，准确率和召回率也没有显著提升。本文提出的互联网药品信息三维语义词典，采用基于本体的语义构造方法和技术，归纳了药品领域的常用术语和概念，综合考虑互联网药品信息对象在文字描述、应用领域、展现形式等多个维度的异构性，从而形成了较完备的推理规则和约束条件。通过具体应用和实验测试，证明了其能有效解决 Web 信息抽取过程中信息不一致的干扰，为准确、自动的抽取提供技术支持，并能显著提高抽取的准确率和召回率。

在互联网药品信息三维语义词典的具体构造上我

们借鉴了 WordNet，采用人工建模和自动扩充相结合的方式。首先针对互联网药品信息对象涉及的概念，结合其在应用领域的通用特征，手工提取语义信息。接着对信息对象的超文本标记和对象属性，使用 XML RDF Schema 进行语义提取，从而形成语义词典的静态部分。为了能不断对语义词典进行扩充，我们设计采用本体集成和关联规则挖掘技术。本体集成技术通过融合外部语义词典，对原语义词典的词汇进行同义词扩充。关联规则挖掘技术是对抽取结果的历史数据进行分析处理，从中得到感兴趣的关联规则，进而利用这些关联规则来补充和优化原语义词典。

在实际建模中我们采用目前使用较广泛的本体描述语言 OWL-Full，因为 OWL-Full 允许使用者定义适合自己使用的属性及其关系，具有很大的灵活度。同时使用开源的 Protégé和 Jena 等工具，来构建和维护语义词典。图 2 是实际编程中基于语义词典转换而来的 XML 片段。

```
<property name="hasUnit" name_cn="计量单位">
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])[商药]\s*品\s*单\s*位(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])单\s*位(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])计\s*量\s*单\s*位(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <value-keywords>盒,瓶,支,客</value-keywords>
</property>

<property name="hasApprovalCode" multiple-value="true" name_cn="批准文号">
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])批\s*准\s*文\s*号(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])生\s*产\s*批\s*号(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])批\s*件\s*信\s*息(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <value-regex>(?:国药(准试|字)?[a-zA-Z]*[0-9]+</value-regex>
    <value-regex>.*字.*\d{4}.*第.*[0-9a-zA-Z]+.*号.*</value-regex>
</property>

<property name="hasManufacturer" name_cn="生产厂商">
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])生\s*产\s*厂\s*[家商](?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <tag-regex>(?:^|[^\u4e00-\u9fa5a-z0-9])生\s*产\s*企\s*业(?:[^\u4e00-\u9fa5a-z0-9]+|$)</tag-regex>
    <value-keywords>公司,厂,药业,制药,Co.,Ltd,Inc,德国,英国,法国,美国,瑞士,加拿大,西班牙</value-keywords>
</property>
```

图 2　互联网药品信息三维语义 XML 片段

3.2 结合网页结构识别信息关联

在对多个互联网医药电子商务网站进行调研后，我们总结出三类包含药品信息的页面：导航推荐页、分类列表页和销售详细页。它们提供药品信息的丰富程度不一，横向呈现依次递减，即一张网页中所包含不同药品的个数逐渐减少；纵向呈现依次递增，即一张网页中同一个药品所罗列不同属性的个数逐渐增加。为保证所抽取到互联网药品信息的全面和准确，目标源网页应锁定药品销售详细页。再者，导航推荐页和分类列表页中出现的每个药品实际都存有指向各自对应销售详细页的 URL，为避免抽取信息重复，需要过滤这两类网页。图 3 是一张典型的药品销售详细页面。

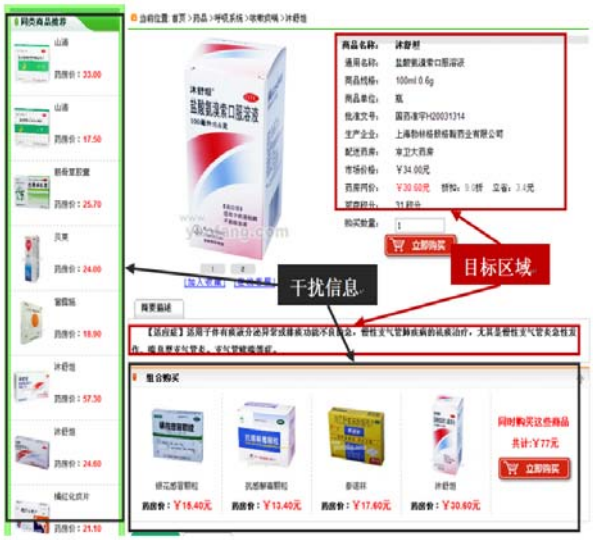


图 3　典型的互联网药品销售详细信息页面

其中目标区域是需要抽取的药品属性信息集，同时页面上还存在很多干扰信息，例如同类商品推荐、组合购买推荐等等，因此在使用互联网药品信息三维语义词典中定义的关键词对页面中目标信息进行定位时，会匹配到很多候选信息，而这些信息之间往往存在着多种不同的关联。我们的抽取目标是找到一组可以充分表征该销售详细页对应的药品属性信息集，这就需要从零乱无序的候选命中片段中找出相互关联的信息并进行结构化处理。

HTML 网页具有一定的层次结构，例如 标记用来排版分块大段网页内容，<ul>、<li>标记可以显示列表结构，规范的表格通常使用<table>进行标记，还有诸如<p>、
、<form>等等。利用这些特征，我们调研了 5 家不同的医药电子商务网站，通过对大量的药品销售详细页面样本进行分析发现，需要抽取的目标药品属性信息往往在网页结构上呈现一定的聚集度，例如同为表格项或列表项。但由于 HTML 代码没有严格的编写格式，其中又经常被嵌入许多脚本语言，例如 JavaScript，这就造成在 HTML 源码字符串中文字的相对位置和距离无法绝对正确地和浏览器中看到的网页结构形成映射关系，因此我们不能简单地基于 HTML 源码字符串来判断其中文字的聚集度。HTML 网页是以 HTML 标签相互嵌套形成的

树形结构文档。经过研究，我们决定将 HTML 网页抽象成 DOM 树的形式，使不同文字节点的逻辑结构关系能得以保留。

DOM(Document Object Model)即文档对象模型，是 W3C 组织推荐的处理可扩展置标语言的标准编程接口。在 Web 信息抽取研究领域，先将网页转换为 DOM 树再进行结构匹配和信息抽取的方法已有一定的应用。早在 1977 年，S.Selkow 就提出了一种名为 Tree Edit Distance[7]的树匹配算法，他认为两棵树之间可以通过增加节点、删除节点和更换节点三种操作完成相互转换，并且对每种节点操作赋予不同的代价，则两颗树的相似程度可以通过计算它们相互转换的最小代价来加以衡量。该算法虽然经过了严密的理论论证，但实际操作的复杂度极高。为了保证抽取算法切实可行，我们根据信息熵的理论，提出了网页结构语义熵[8]的概念和计算方法来量化表示目标命中信息集的聚集度，替代原有的树匹配算法，同样能够识别目标抽取信息的结构模式。

3.3 抽取算法整体设计思路

3.3.1 目标源网页的获取

首先，使用元搜索技术，通过配置好的搜索引擎、比价网站等，输入可能表征医药电子商务网站的关键词组合，例如“网上药店”、“药品名称 生产厂家”、“批准文号 价格”等具有一定语义关联的领域词汇，从搜索结果页中获得一定数量的 URL，取出其网站顶级域名组成一个待访问网站集合。然后通过网络爬虫遍历每个候选网站，取得所有的站内网页作为抽取算法进一步分析的目标源网页。

3.3.2 网页 DOM 树的生成

由于 HTML 文档的结构定义较为松散，目前互联网上 HTML 文档常常都不完全符合 HTML 规范，甚至有时候会出现诸如遗漏结束标签等错误。而绝大多数浏览器的解析引擎都包含了对 HTML 文档进行纠错的功能，所以在网页展现时通常都能保持正常的显示效果。但是，在分析网页结构的过程中，错乱的标签嵌套会对网页的树形结构有所影响，进而导致网页结构语义熵计算错误，所以在抽取前要对 HTML 文档进行预处理，修正错误的标签并过滤一些格式控制标签。在实际编程中我们使用开源的 NekoHTML 包来完成上述预处理工作，生成网页 DOM 树。图 3 中目标区域的 HTML 代码片段经过处理可得图 4 所示的 DOM 树结构。

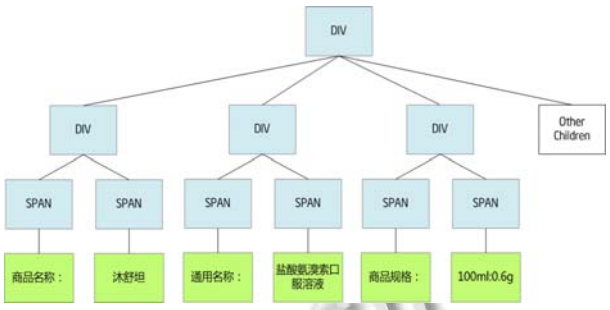


图 4 网页 DOM 树结构示例

3.3.3 网页结构语义熵的计算

信息熵是用来表征一个系统的平均信息量。同样，我们提出的结构语义熵概念是用来表征一张网页所包含特定领域知识的平均信息量。由于我们将网页转化成 DOM 树的形式，强调了网页的层次结构，并增加了对网页内容的语义关联约束，故命名为结构语义熵，具体定义如下：

设 DOM 树中的节点 S 包含 s 个叶节点，每个叶节点可按提供目标属性信息的种类划分成 m 个不同的类别 C_i，$i \in \{1,2,3,,m\}$，其中未提供目标信息的节点归为同一类，设 s_i 为属于类别 C_i 的叶节点的个数，用以下公式计算 DOM 树中任一节点的结构语义熵：

$$E(S) = -\sum_{i=1}^{m} p_i \log(p_i) \quad (1)$$

其中 p_i 是任意一个叶节点属于类别 C_i 的概率，可以按 s_i / s 计算，log 函数以 2 为底，因为按信息论原理，信息都是按位进行编码的[8]。通过对大量网页 DOM 树样本的计算实验发现，一个节点的结构语义熵可以用来衡量该节点下目标领域语义的丰富程度，抽取算法应选择结构语义熵最大的节点作为信息聚集节点，列表节点(对应推荐药品、分类药品等噪声信息)的结构语义熵一般略大于 1。

利用之前使用语义词典匹配到的所有包含药品属性的文本节点，可以对 DOM 树中的各节点进行结构语义熵的计算，取出其中高于指定阈值且值最大的节点作为目标区域子树的根节点，进而抽取该子树所包含的目标信息。

3.3.4 信息抽取

在之前语义词典标注的基础上，对于已命中的药

品属性节点可直接抽取节点内的文本内容进行分析。如果除去属性名称外并未抽取到对应的属性值文本，则属性值和属性名很可能在网页 DOM 树上分别位于两个兄弟节点，例如图 4 的情况，已命中的节点包括“商品名称：”、“通用名称：”、“商品规格：”等，在各自对应的兄弟节点上我们可以抽取到“沐舒坦”、“盐酸氨溴索口服溶液”、“100ml:0.6g”等属性值。

对于语义词典中未标注的属性信息，由于网页内容自上而下、从左到右的展现顺序和网页 DOM 树的深度优先遍历相对应，故可根据网页的逻辑结构，使用启发式的方法先进行抽取，再通过某些药品属性取值的特征规律进行筛选。

4　抽取算法验证实验

本实验的提出一方面为了验证我们的互联网药品信息抽取算法流程是否能正确执行，另一方面为了统计算法抽取目标药品属性信息的准确率和召回率，用以分析算法是否能达到全面准确实时自动抽取互联网药品信息的要求。

4.1 实验验证方法

如何判断通过计算网页结构语义熵定位到的目标节点包含我们在网页上需要抽取的药品属性信息？人工验证当然是最可靠的方式，但实验样本数量巨大，必须考虑使用程序自动判断。根据网页 DOM 树模型，我们确立了程序验证的关键点，即对同一张网页的 DOM 树结构，使用假定为准确的目标信息节点定位程序输出结果节点集，再和结构语义熵定位程序输出的节点集进行匹配，判断各节点是否在原网页 DOM 树上对应同一个节点，进而统计出结构语义熵判断和抽取目标药品属性信息的准确率和召回率。

目前已经有不少 DOM 树节点定位的方式算法，例如 WebOQL[9]使用常见的数据库 select-from-where 架构来查询网页 DOM 树上特定信息的节点，W4F[10]则使用 XPath 进行节点定位。经过研究我们发现，在 Web 信息抽取领域应用广泛的基于正则表达式的信息匹配抽取方法，其本质和 XPath 很类似，都是根据既定的规则模板在网页全文中找出所有匹配模板的部分，区别仅仅在于正则表达式是基于网页源码字符串的，而 XPath 则是应用在网页 DOM 树结构之上。正则表达式的规则描述的是句法，XPath 的规则描述的是路径，但两者在匹配过程中都利用了丰富的 HTML 标签作为判断根据。我们决定基于 XPath 抽取模板编写准确的 DOM 树节点定位程序作为结构语义熵算法的参照物。

4.2 实验数据准备

我们选取 818 上海药房网(www.818shyf.com)、百洋健康药房(www.baiyjk.com)、百姓药房网(www.bxdyf.com)、惠好连锁网(www.511yd.com)、药房网(www.yaofang.cn)作为测试目标网站。通过对这 5 个网站进行调研，我们发现其销售的药品种类数量均达到了一定数量规模，能为实验提供大量测试网页样本。同时，在需要抽取目标信息的药品销售详细页中，存在着大量的其他药品广告或相关药品推荐列表，此外这些网站还按照药品分类导航提供了非常多的药品销售列表页面，这些噪声信息都能有效检验我们抽取算法的稳定性和普适性。

4.3 实验结果分析

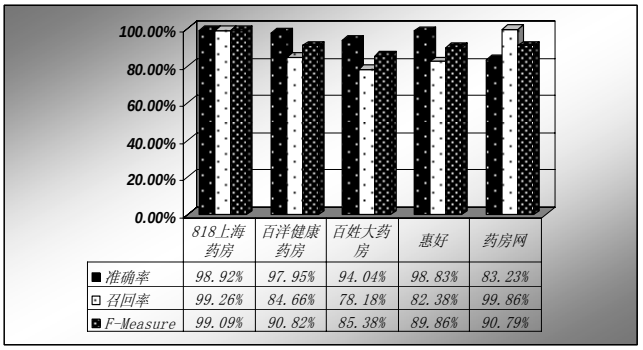
本实验中召回率、准确率和 F-measure 的计算方法如下：

(1) 召回率 R：页面内抽取正确的药品属性值条数 / 页面内实际包含的药品属性值条数；

(2) 准确率 P：页面内抽取正确的药品属性值条数 / 页面内抽取到的药品属性值条数；

(3) F-measure：信息检索术语，是召回率与准确率的综合指数。这里为 2RP / (R + P)。

实验基于结构语义熵阈值为 1 进行，总共分析了 5 个网站共计 37656 张网页，算法流程执行正确，目标信息抽取平均 F-measure 为 91.2%。图 5 按照不同网站药品属性名值对的抽取结果进行横向统计：

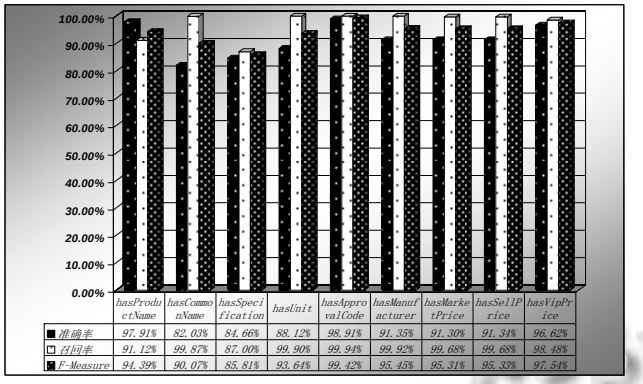


	818上海药房	百洋健康药房	百姓大药房	惠好	药房网
准确率	98.92%	97.95%	94.04%	98.83%	83.23%
召回率	99.26%	84.66%	78.18%	82.38%	99.86%
F-Measure	99.09%	90.82%	85.38%	89.86%	90.79%

图 5　按网站分别统计药品属性名值对抽取结果

从图中易见，我们的抽取算法准确率和召回率较高。其中药房网准确率偏低的原因是该网站当时的版

本存在着多个地区分站，和主站的网页风格不同，XPath 模板只是根据其主站进行编写，因而在分析分站网页时出现失配次数多。图 6 节选了部分抽取结果较多的药品属性进行纵向统计：



	hasProductName	hasCommonName	hasSpecification	hasUnit	hasApprovalCode	hasManufacturer	hasMarketPrice	hasSellPrice	hasVipPrice
准确率	97.91%	82.03%	84.66%	88.12%	98.91%	91.35%	91.30%	91.34%	96.62%
召回率	91.12%	99.87%	87.00%	99.90%	99.94%	99.92%	99.68%	99.68%	98.48%
F-Measure	94.39%	90.07%	85.81%	93.64%	99.42%	95.45%	95.31%	95.33%	97.54%

图 6 按目标属性分别统计药品属性名值对抽取结果

抽取结果同样证明了该抽取算法准确率和召回率较高。商品名称(hasProductName)的召回率明显低于其他属性，原因在于像百姓大药房这类网站，其商品名称前没有冠以类似“商品名称”的属性标识，我们的抽取算法目前尚不支持无属性标识节点的匹配。

5 结论与展望

现有的 Web 信息抽取技术在准确率、召回率和人工干预程度方面都存在着各自的局限和不足，无法满足政府部门全面准确地对互联网药品信息进行抽取和监测的需求。本文提出并设计了一种新的互联网药品信息抽取算法，引入语义技术作为突破口，通过构建互联网药品信息三维语义词典，屏蔽不同医药电子商务网站在网页内容和结构上的异构性，同时利用药品信息网页中所需抽取的目标药品属性信息具有一定聚集度的特征，基于信息熵的基本理论对目标药品属性信息进行智能定位和抽取。文章首先阐述了国内外在 Web 信息抽取领域的研究现状，对常用技术分别进行了优缺点分析，提出了应用语义技术实现智能抽取的研究方向。接着介绍了本文抽取算法的总体思路，并对其中的关键技术——互联网药品信息三维语义词典和网页结构语义熵进行了详细阐述。最后通过完整的实验证明了该抽取算法不仅可行性高、人工干预少，同时保证了较高的抽取准确率和召回率。

虽然该抽取算法已能在一定范围内智能识别不同风格的药品销售详细页面并自动定位抽取目标药品属性信息，但目前仍无法完全规避不同网站发布药品信息时的不一致。例如算法存在对某些没有属性名称标识的药品属性暂时无法识别等问题，还有待进一步的研究。

参考文献

1 Kushmerick N, Weld D, Doorenbos R. Wrapper Induction for Information Extraction. 15th Int’l Conf. Artificial Intelligence (IJCAI). 1997. 729－735.

2 Muslea I, Minton S, Knoblock C. A Hierarchical Approach to Wrapper Induction. Third Int’l Conf. Autonomous Agents (AA ’99). 1999.

3 Crescenzi V, Mecca G, Merialdo P. RoadRunner: Towards -Automatic Data Extraction from Large Web Sites. 26th Int’l Conf. Very Large Database Systems (VLDB). 2001. 109－118.

4 Chang CH, Lui SC. IEPAD: Information Extraction Based on Pattern Discovery. 10th Int’l Conf. World Wide Web(WWW). 2001. 223－231.

5 Soderland S. Learning Information Extraction Rules for Semi -Structured and Free Text. Journal of the Machine Learning, 1999,34(1-3):233－272.

6 Freitag D. Information Extraction from HTML: Application of a General Learning Approach. 15th Conf. Artificial Intelligence(AAAI ’98). 1998.

7 Selkow S. The Tree-to-Tree Editing Problem. Journal of the Information Processing Letters, 1977:184－186.

8 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.

9 Arocena GO, Mendelzon AO. WebOQL: Restructuring Documents, Databases, and Webs. 14th IEEE Int’l Conf. Data Eng. (ICDE). 1998,24－33.

10 Saiiuguet A, Azavant F. Building Intelligent Web Applications Using Lightweight Wrappers. Journal of the Data and Knowledge Eng., 2001,36(3):283－316.