# 电商网页中商品规格信息自动抽取方法研究

赵晓永,王 磊
ZHAO Xiaoyong, WANG Lei

北京信息科技大学 信息管理学院,北京 100129
School of Information and Management, Beijing Information Science & Technology University, Beijing 100129, China

**ZHAO Xiaoyong, WANG Lei. Product specification auto extract method of e-commerce websites. Computer Engineering and Applications, 2017, 53(24):168-171.**

**Abstract:** The automatic mining of billions of product specification information in Web has important application value in many fields such as e-commerce market analysis, commodity recommendation, after-sales service and so on. But the current methods of specification extraction don't effectively solve the balance between manual annotation workload, scalability and accuracy. This paper proposes the Title Seed Automatic Extract(TSAE) method, using unsupervised learning method, using the page title as seed, combining with statistical characteristics, natural and machine semantics, it achieves higher accuracy while reducing the workload, enhancing the scalability. The experimental results show that the TSAE method has better automatic extraction precision while providing good performance and expansibility, can support the massive data processing, has good practical value.

**Key words:** information extraction; automatic extraction; product specification; e-commerce

**摘 要:** Web中数十亿的商品规格信息的自动挖掘,对电子商务领域的市场分析、商品推荐、售后服务等诸多领域有重要的应用价值。但目前的商品规格信息抽取方法尚未有效解决人工标注工作量、扩展性和准确率之间的平衡问题,提出一种商品网页规格信息自动抽取方法TSAE(Title Seed Automatic Extract),采用无监督的学习方法,以网页标题为种子,结合统计特征、自然语义和机器语义,在减少工作量、提升扩展性的同时,达到了较高的准确率。实验表明,TSAE方法在提供更好的自动化抽取效果的同时,具备良好的性能和扩展性,能够支撑海量数据处理,具有良好的实用价值。

**关键词:** 信息抽取;自动抽取;商品规格信息;电子商务
**文献标志码:** A **中图分类号:** TP311 doi:10.3778/j.issn.1002-8331.1708-0053

## 1 引言

Web作为海量的结构化数据源,其中数十亿的商品属性规格信息的自动挖掘,对市场分析、商品推荐、售后服务等诸多领域有重要的应用价值[1]。

商品规格信息,也称商品属性信息,包含足以反映商品品质的主要指标,如成分、含量、容量、长短、粗细等,该信息的抽取作为文本信息抽取的一个重要子问题[2],已得到了较多的研究,从训练方法角度可分为监督/半监督和无监督两类。监督/半监督方法需要人工构造本体库、领域词典,或标记部分属性值和属性名等,准确度高,但工作量大、通用性和扩展性较差。其中,文献[3]以DOM树解析获得的文本节点为抽取对象,利用条件随机场CRF模型提取拍卖网站中描述商品规格属性的文字信息,但并没有进一步抽取商品属性名、属性值及对应关系;文献[4]提出了一种基于领域本体并结合视觉信息的方法从网页表格中获取商品"属性-值"关系,但首先需要人工构建领域本体,极大降低了方法的移植性;文献[5]基于隐马尔可夫模型(HMM)进行信息抽取,应用模糊积分单调性建立基于Choquet积分的隐马尔可夫模型(CI-HMM),解决了HMM观察序列概率计算所需的条件独立性假设,优化了HMM观察序列的计算,并以网上书店商品数据实证表明,模型有良好的

**基金项目:** 国家自然科学基金(No.61572079);北京市教育委员会科技计划一般项目(No.KM201711232018)。
**作者简介:** 赵晓永(1981—),男,博士,研究方向为Web数据挖掘和大数据方向,E-mail:zhaoxiaoyong@bistu.edu.cn;王磊(1982—),女,博士,研究方向为服务计算和大数据方向。
**收稿日期:** 2017-08-04 **修回日期:** 2017-10-11 **文章编号:** 1002-8331(2017)24-0168-04

适用性和精确度；文献[6]提出一种有监督的大规模商品规格信息发现与抽取方法DEXTER，首先利用监督学习方法识别出网页中商品规格信息所在的HTML片段，然后分别采用领域无关的无监督包裹器生成方法和文献[7]中提出的能容忍噪声的包裹器生成方法两种轻量级策略从HTML片段中抽取属性值对。

无监督方法不需要人工参与，通用性和扩展性更好，但准确性较低。其中，文献[8]提出一种基于网页标题的无监督模板构建方法，首先利用商品网页标题构建领域相关的属性词包，然后基于预设分隔符细化文本节点，结合领域商品属性词包获取种子“属性-值”关系，最后结合网页布局信息和字符信息来筛选与构建模板。不过该方法未充分利用网页结构中的语义信息，准确率较低。文献[9]提出一种无监督的中文商品属性结构化方法，借助搜索引擎，基于小概率事件原理分析文法关系来抽取属性值与属性名，同时提出相对不选取条件概率场，并使用PageRank算法来计算属性值与属性名的配对概率。不过该方法只抽取商品标题中出现的属性值，抽取到的信息完整度较低。

本文提出了一种结合统计特征、自然语言语义与HTML语义标签的商品规格数据自动抽取方法TSAE（Title SeedAutomatic Extract）。在启发式规则基础上，首先以网页标题为种子，基于商品网页标题的自然语义，对其分词后构建领域相关的属性值特征词典，结合该词典、网页统计特征和HTML语义标签自动识别出商品属性值信息所在的多个区域，然后在这些区域内，遍历语义标签和表格标签，并结合自然语义分隔符（通常为冒号），自动抽取出属性-值对信息。最后，分别使用英文的商品特征抽取金标准数据集[10]和爬虫抓取的中文电商网站商品数据进行实验，结果表明本文提出的方法在准确率和召回率方面都有了较为明显的提升，同时具备良好的时间和空间效率，能够适应海量数据处理的要求。

## 2　方法设计

### 2.1　启发式规则

通过文献[6-8，11]的总结和对主要电子商务网站[12]的分析，发现有如下启发式规则：

**规则1**　商品网页标题中通常包含了商品规格信息中的部分属性值和前后缀（网站名称、域名等）。

这利于提高商品被搜索引擎收录和检索的几率，如：“【飞利浦55PUF6092/T3】飞利浦(PHILIPS)55PUF6092/T3　55英寸　64位九核4K超高清智能液晶平板电视机（银灰色京东微联APP控制）【行情　报价　价格　评测】-京东”包含了商品规格信息中的产品品牌、屏幕尺寸和屏幕分辨率三个属性的值，如图1所示。


**图1　京东商品规格信息示例**

**规则2**　商品的属性信息通常采用HTML文本语义标签[dl、dd、dt]、[ul、li]和表格标签[th，td]、[td，td]，且商品属性区域内主要为表格子节点（tbody、tr、th、td等）、文本内容子节点（dl、dd、dt、ul、li等）和内联文本语义子节点（i、em、b、u、strong、sub等），其他类型的子节点（img、video、a、input、button等）所占比例较低。

这也便于浏览器布局和搜索引擎优化，如表1。

**表1　网站对属性信息采用的语义标签**

| 标签类型 | 采用的网站 |
|---|---|
| [ul、li] | 淘宝、天猫、苏宁、bestbuy等 |
| [dl、dd、dt] | 京东、一号店、newegg等 |
| [th,td]、[td，td] | amazon、ebay、唯品会等 |

**规则3**　属性名和属性值位于文档对象模型（Document Object Model，DOM）中同一个文本节点内时，属性名称文本后通常带有自然语言的分隔符，通常为冒号。

### 2.2　整体流程

在启发式规则指导下，本文提出的商品网页属性-值对规格数据自动抽取方法整体流程如图2所示。

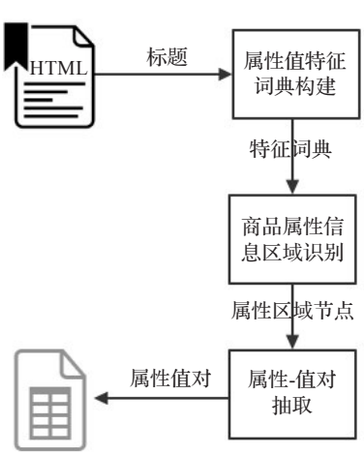


**图2　TSAE整体流程图**

首先对网页标题预处理并分词后构建出领域相关的属性值特征词典，然后结合该特征词典、网页统计特征和HTML语义标签识别出商品属性值信息所在的多个区域，最后在这些区域内，遍历语义标签和表格标签，并结合自然语义分隔符（通常为冒号），自动抽取出属性-值对信息。

### 2.3 构建属性值特征词典

为提高商品被搜索引擎收录和检索的几率,基于启发规则1,商品网页标题中通常包含了商品销售属性值和前后缀(网站名称、域名等),通过分词可将标题转化为属性值,但存在两个问题:

(1)网站名称在网页内会多次出现,如果将其加入属性特征词典内,将对后续分析增加噪声。

(2)对中文电商网站而言,标题文本较短,其中包含了品牌名称、规格等大量新词和参数等量词,对分词方法的未登录词识别能力要求高。

条件随机场(ConditionalRandom Fields,CRF)[13]是由Lafferty等人于2001年提出的一种用来标记和切分序列化数据的统计模型,其模型思想的主要来源是最大熵模型,具有特征选择灵活和拟合程度好等优点,且不存在标记偏置问题。基于CRF的中文分词方法[14],可有效解决交叠歧义和未登录词带来的不确定性,提高品牌名称、规格等大量新词识别能力和分词系统的准确率与召回率。因此,本文选择基于CRF模型的分词方法对中文标题分词。

算法1描述了构建属性值特征词典的具体过程。

**算法** 1 属性值特征词典构建算法

输入:网页内容 $S$ 。

输出:属性值特征词典 $D$ 。

(1)获取网页title标签中内容,作为标题 $t$ 。

(2)使用最长公因子匹配算法,在网页去除title标签后的内容中进行比对,将结果作为新的标题 $t$ 。

(3)使用正则表达式/[\u4E00-\u9FA5\uF900-\uFA2D]+/识别是否为中文标题,如果是,继续下一步骤(4);否则使用空白分隔符切词后加入 $D$ 中,转到步骤(6)。

(4)使用CRF模型对标题 $t$ 进行分词,并加上词性标注。

(5)将分词结果的名词、形容词、数量词词性[15]的序列项及新词作为属性值特征词典 $D$ 。

(6)输出属性值特征词典 $D$ ,算法结束。

### 2.4 识别商品属性信息区域

商品属性值特征词典描述了商品销售属性词,基于启发规则2,网页中商品属性信息的描述中部分或全部包含了这些特征词,算法2结合HTML语义标签和文献[11]中的统计规则,自动识别出商品属性值信息所在区域的HTML片段,具体过程如下。

**算法** 2 商品属性信息区域识别算法

输入:商品属性特征词典 $D$、网页内容 $S$ 。

输出:商品属性信息所在区域根节点DOM集合 $N$ 。

(1)针对 $D$ 中的每个词,从 $S$ 中检索该词,如果找到匹配的词,且该词所在的DOM节点 $e$ 不在候选集合,则将 $e$ 加入到候选集合 $E$ 中。

(2)对候选集合 $E$ 中的每个节点 $e$ ,重复步骤(3)~(5)。

(3)查找最近的上级 $[td、dd、li]$ 类型父节点 $E_{tdl}$ ,如未找到,则排除该候选节点 $e$ ;否则,继续下一步骤。

(4)查找 $E_{tdl}$ 分别对应的上一级 $[table、dl、ul]$ 类型的父节点 $E_{tdu}$ ,如未找到,则排除该候选节点 $e$ ;否则,从 $E_{tdl}$ 节点集合中删除属于该 $E_{tdu}$ 的子节点,继续下一步骤。

(5)在文献[11]的统计规则基础上,结合启发式规则2,计算 $E_{tdu}$ 的表格子节点 $[tbody、tr、th、td]$、文本内容子节点 $[dl、dd、dt、ul、li]$ 和内联文本语义子节点 $[i、em、b、u、strong、sub]$ 的数量所占比率 $r$ ,如果 $r$ >阈值 $t$ ( $t$ 默认80%),则将节点 $e$ 加入结果集合 $N$ 中。

(6)输出结果集合 $N$ ,算法结束。

### 2.5 抽取属性-值对

对算法2计算出的候选商品属性信息区域DOM集合 $N$ ,重复执行以下算法进行属性-值对关系的抽取。

**算法** 3 属性-值对关系抽取算法

输入:商品属性信息区域DOM集合 $N$ 。

输出:属性-值对关系DOM元素字典 $E_{kv}$ 。

(1)筛选出 $N$ 中所有 $dl$ 类型节点 $E_{dl}$ ,如果 $E_{dl}$ 不为空,则对 $E_{dl}$ 的每个元素 $e$ ,重复步骤(2)。

(2)遍历 $e$ 的 $dt$ 类型子节点 $e_{dt}$ ,对每个节点,找到其下一个 $dd$ 类型兄弟节点 $e_{dd}$ ,将 $(e_{dt}, e_{dd})$ 加入 $E_{kv}$ 。

(3)筛选出 $N$ 中所有 $tr$ 类型节点 $E_{tr}$ ,如果 $E_{tr}$ 不为空,则对 $E_{tr}$ 的每个元素 $e_x$ ,重复步骤(4)、(5)。

(4)筛选出 $e_x$ 的 $th$ 类型子节点 $e_{th}$ ,如果 $e_{th}$ 不为空,则对每个节点,找到其下一个 $td$ 类型兄弟节点 $e_{td}$ ,将 $(e_{th}, e_{td})$ 加入 $E_{kv}$ 。

(5)如果 $e_{th}$ 为空,则遍历 $e_x$ 的 $td$ 类型子节点 $e_{td}$ ,对每个节点,找到其下一个 $td$ 类型兄弟节点 $e_{std}$ ,将 $(e_{td}, e_{std})$ 加入 $E_{kv}$ ,并跳过 $e_{std}$ 。

(6)遍历 $e_x$ 的 $li$ 类型子节点 $e_{li}$ ,对每个节点 $e_y$ ,重复步骤(7)、(8)。

(7)如果 $e_y$ 只包括一个(文本类型)子节点,基于启发规则3,如果该子节点的text值匹配正则模式/^|S+[:]|s*|S+/g,则将该 $li$ 节点 $(e_y, e_y)$ 加入 $E_{kv}$ ,否则继续下一个节点;如果 $e_y$ 有多个子节点,则继续下一步。

(8)遍历 $e_y$ 的下一级子节点 $e_{cli}$ (包括文本类型节点),对 $e_{cli}$ 中每个节点 $e_z$ ,找到其下一个text值不为空白的兄弟节点 $e_v$ ,则将 $(e_z, e_v)$ 加入 $E_{kv}$ 。

(9)输出 $E_{kv}$ ,算法结束。

## 3 实验

### 3.1 实验环境

本文的实验环境为1台8核16 GB内存,1 TB硬盘的DELL R630服务器,操作系统为CentOS 7.0_x64。

测试数据分为中、英文两类。其中,英文测试数据

采用商品特征抽取金标准数据集[10]，该数据集包含32个电商网站的商品网页数据。中文测试数据使用爬虫从排名前10[12]的中文电商网站爬取，每个网站20个商品网页，共200个商品网页，并人工标注出了商品属性信息。

采用文献[6]中提出的DEXTER、文献[11]中提出的方法和文献[10]中提出的字典法作为基准方法，与本文提出的方法进行对比。

### 3.2　实验结果

表2显示了本文提出的属性信息区域识别算法与其他方法的对比结果（字典法中无此步骤，无法对比），其中TSAE-金标准为TSAE算法针对金标准数据集的结果，TSAE-中文为TSAE算法针对爬取数据的结果。可以看出，本文方法在两种数据集上的 $F$ 值都有了较为显著的提升。

**表2　属性信息区域识别结果对比**

| 方法 | 准确率 | 召回率 | $F1$值 |
|---|---|---|---|
| DEXTER | 0.721 | 0.746 | 0.733 |
| 文献[11]方法 | 0.764 | 0.752 | 0.758 |
| TSAE-金标准 | 0.824 | 0.843 | 0.833 |
| TSAE-中文 | 0.912 | 0.864 | 0.887 |

表2结果与启发式规则1的观察一致，网页标题中包含的部分商品属性信息可用于更准确地定位到商品属性信息区域，从而提高了这些区域的识别准确率。

表3显示了本文的TSAE算法与其他方法抽取商品属性的对比结果，可以看出，得益于商品属性信息区域识别效果的提升，本文方法的抽取准确率和 $F$ 值均有了较为显著的提升。

**表3　抽取结果对比**

| 方法 | 准确率 | 召回率 | $F1$值 |
|---|---|---|---|
| 字典法 | 0.487 | 0.568 | 0.524 |
| DEXTER | 0.619 | 0.701 | 0.658 |
| 文献[11]方法 | 0.751 | 0.691 | 0.719 |
| TSAE-金标准 | 0.817 | 0.789 | 0.803 |
| TSAE-中文 | 0.901 | 0.827 | 0.862 |

## 4　结论

本文针对目前的商品属性规格信息抽取方法尚未有效解决人工标注工作量、扩展性和准确率之间的有效平衡问题，结合统计特征、自然语义和机器语义，提出一种商品网页属性-值对数据自动抽取方法TSAE，采用无监督的学习方法，在减少工作量、提升扩展性的同时，达到了较高的准确率。实验表明，TSAE方法在提供更好的自动化抽取效果的同时，具备良好的性能和扩展性，能够支撑海量数据处理要求，具有良好的实用价值。不过，标题中通常只包含属性值的关键词描述，而本体则包含了对属性名的规范化描述，如何与本体应用结合，提升自动抽取的准确性和规范化，是需要进一步研究的问题。

## 参考文献：


[1] Huang J M，Wang H X，Jia Y，et al.Link-based hidden attribute discovery for objects on Web[C]//Proceedings of the 14th International Conference on Extending Database Technology，2011：473-484.

[2] Ghani R，Probst K，Liu Y，et al.Text mining for product attribute extraction[J].ACM SIGKDD Explorations Newsletter，2006，8(1)：41-48.

[3] Wong T L，Lam W.Adapting Web information extraction knowledge via mining site-invariant and site-dependent features[J].ACM Transactions on Internet Technology，2007，7(1)：6.

[4] Holzinger W，Krüpl B，Herzog M.Using ontologies for extracting product features from Web pages[C]//International Conference on the Semantic Web.[S.l.]：Springer-Verlag，2006：286-299.

[5] 邓斌，邵培基，夏国恩.基于Choquet积分的HMM商品信息抽取方法[J].系统工程，2008(12)：110-114.

[6] Qiu D，Barbosa L，Dong X L，et al.DEXTER：Large-scale discovery and extraction of product specifications on the web[J].Proceedings of the VLDB Endowment，2015，8(13)：2194-2205.

[7] Dalvi N，Kumar R，Soliman M.Automatic wrappers for large scale web extraction[J].Proceedings of the VLDB Endowment，2011，4(4)：219-230.

[8] 唐伟，洪宇，冯艳卉，等.网页中商品“属性-值”关系的自动抽取方法研究[J].中文信息学报，2013，27(1)：21-29.

[9] 侯博议，陈群，杨婧颖，等.无监督的中文商品属性结构化方法[J].软件学报，2017，28(2)：262-277.

[10] Petrovski P，Primpeli A，Meusel R，et al.The WDC gold standards for product feature extraction and product matching[M].[S.l.]：Springer International Publishing，Cham，2017：73-86.

[11] Petrovski P，Bizer C.Extracting attribute-value pairs from product specifications on the Web[C]//International Conference on Web Intelligence(WI'17)，Leipzig，Germany，2017：558-565.

[12] Alexa.Top sites by category：Shopping[EB/OL].(2017-07).http://www.alexa.com/topsites/category/Shopping.

[13] Lafferty J，McCallum A，Pereira F.Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C]//International Conference on Machine Learning，2001.

[14] 刘泽文，丁冬，李春文.基于条件随机场的中文短文本分词方法[J].清华大学学报：自然科学版，2015(8)：906-910.

[15] 刘群，张华平，张浩.计算所汉语词性标记集[EB/OL].(2012-01)[2017-07].http://ictclas.nlpir.org/nlpir/html/readme.htm.