文章编号：1003-0077(2013)01-0021-09

网页中商品“属性—值”关系的自动抽取方法研究

唐 伟，洪 宇，冯艳卉，姚建民，朱巧明

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：商品属性及其对应值的自动挖掘，对于基于 Web 的商品市场需求分析、商品推荐、售后服务等诸多领域有重要的应用价值。该文提出一种基于网页标题的模板构建方法，从结构化网页中抽取完整的商品“属性—值”关系。该方法包含四个关键技术：1）利用商品网页标题构建领域相关的属性词包；2）基于预设分隔符细化文本节点；3）结合领域商品属性词包获取种子“属性—值”关系；4）结合网页布局信息和字符信息来筛选与构建模板。该文的实验基于相机和手机两个领域展开，获得 94.68％的准确率和 90.57％的召回率。

关键词：商品“属性—值”关系抽取；Web 数据挖掘；模板构建

中图分类号：TP391　　**文献标识码**：A

Automatic Extraction of the Product “Attribute-Value” Pair from the Webpages

TANG Wei, HONG Yu, FENG Yanhui, YAO Jianmin, ZHU Qiaoming

(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: If we represent the products as attributes and attribute values, it will improve the effectiveness of many applications, such as demand forecasting, product recommendations, and product supplier selection. In this paper, we propose a novel pattern based method to extract the “attribute-value” pair of product from structured or semi-structured Web pages. This approach contains four key components: 1) acquire domain-specific attributes from titles of Web pages in the same domain. 2) refine text nodes based on some default delimiters. 3) collect seed “attribute-value” pairs based on the domain-specific attributes. 4) construct high-quality patterns by combining page-specific layout information and character information. The experimental corpus is collected from two domains: digital camera and mobile phone. Experiments show the proposed method can schieve 94.68% in precision and 90.57% in recall.

Key words: product “attribute-value” relation extraction; web data mining; template construction

1 引言

互联网技术的高速发展使得在线购物极大普及，国内以淘宝、拍拍为代表的 C2C 网站发展极为迅猛。截至 2010 年底，C2C 电子商务网站中存在着超过 8 亿种商品，而且新的商品仍在不断加入。

通常人们认为商品只是与少数特征（如品牌、尺寸、颜色等）相关联的实体，而忽视了商品详细信息的重要性。把商品描述成一个“属性—值”集合，有利于商品市场需求分析、商品推荐、商品比较、售后服务等诸多应用。本文将描述商品属性名与属性值的对应文本定义为商品的“属性—值”关系。C2C 电子商务平台为抽取完备商品“属性—值”关系提供了数据来源。

根据大量观察发现，卖家为吸引顾客，倾向于将商品包装上详细的商品属性信息以结构化或半结构化的形式展示在网页上，并且同一个网页中商品“属性—值”关系以相似或者相同的方式重复出现，如图 1 中数码相机的“属性—值”关系，即“感光元件—

收稿日期：2011-09-30　**定稿日期**：2012-05-07

基金项目：国家自然科学基金资助项目(60970057)；国家自然科学基金资助项目(61003152)；苏州市自然科学基金资助项目(SYG201030)

作者简介：唐伟(1988—)，男，硕士研究生，主要研究方向为文本挖掘，信息抽取；洪宇(1978—)，男，副教授，主要研究方向为信息检索、自然语言处理。冯艳卉(1987—)，女，硕士研究生，主要研究方向为自然语言处理、网络文本挖掘。

CCD传感器”、“传感器尺寸—1/2.3英寸CCD”等构成的类表格形式。这一现象为利用构建模板的方法抽取商品“属性—值”关系提供了依据。由于不同卖家对于商品信息的组织存在很大的差异，使得连续且高效地从批量网页中抽取商品“属性—值”关系具有挑战性。

商品信息抽取作为文本信息抽取的一个重要子问题[1]，已有很多的相关的研究，如文献[2]以DOM树解析获得的文本节点为抽取对象，利用CRF模型提取拍卖网站中描述商品属性的文字信息。其中文本节点被定义为网页源码中连续的商品信息描述文本片段。以示例网页的网页源码片段“<tr><td>光学变焦</td><td>4倍</td></tr>”为例，其中“光学变焦”和“4倍”就是两个独立但相邻的文本节点。但文献[1-2]并没有获取独立的商品属性名、属性值以及对应关系。文献[3]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。但需要人工构建领域本体，大大降低了方法的移植性。本文通过对淘宝、拍拍等电子商务网站的统计，发现将商品“属性—值”关系组织到同一文本节点中的网页占较大比例，如手机领域的统计结果为24.33%。但文献[3]中默认网页中商品“属性—值”关系组织在不同的文本节点，因此不能处理这类网页。

图1 来自拍拍网的示例网页快照

本文专注于从电子商务网站中获取独立的商品属性名、属性值以及对应关系，针对以上工作的不足，提出一种无监督的模板自动构建方法。该模板融合了网页结构信息和字符信息，能够有效获取描述商品属性文本，识别并抽取出相应的数值、规格和范畴等属性值信息。基于自动构建的领域属性词包，本文的主要工作包括以下方面。

1）细化网页文本节点，统一商品属性名、值位于相同和不同文本节点的解决方法。

2）基于领域商品属性词包，获取每个网页内潜在的属性文本片段（即含有主题词）。

3）通过对属性文本片段的过滤、双向扩展等操作获得种子“属性—值”关系。

4）自动学习种子“属性—值”关系的网页物理结构，并结合字符特征构建候选模板。

5）引入加权词表和筛选阈值，获得能够抽取其他潜在“属性—值”关系的优质模板。

相比于前人的工作，本文主要贡献如下：

1）通过对文本节点的细分，使商品属性名、值处于相同和不同文本节点的解决方法得到统一。而相关工作中并未考虑从同一文本节点中获取分离的商品“属性—值”关系。

2）首次提出利用易获取的领域商品网页的标题自动构建领域属性词包，进而定位、筛选、扩展获得种子“属性—值”关系。此方法避免了人工标注领域知识，具有很好的可移植性。

3）以种子“属性—值”关系为提取单元，融合网页结构信息和字符信息自动构建模板，实现了商品“属性—值”关系的精准抽取。而相关工作中仅仅考虑了网页结构信息，所得到的模板适用性较差。

4）本文在 B2C、C2C 以及大量科技门户网站进行了充分的实验，均取得极佳的效果。

本文的组织结构如下：第 2 节介绍相关工作；第 3 节给出方法框架，并详细介绍其中基于主题特征的候选关系挖掘、模板的筛选与构建以及基于模板的商品“属性—值”关系提取；第 4 节介绍实验设计并分析实验结果；第 5 节总结全文。

2　相关工作

本文主要探讨如何自动地从在线商品销售网页中提取商品“属性—值”关系的问题，本节侧重回顾与这一问题相关的现有研究并给予分析。

文献[3]讨论了基于表格本体和商品属性本体如何实现网页中商品“属性—值”关系的提取。该方法中利用表格本体和网页视图信息识别出网页中潜在的表格结构部分，然后利用商品属性本体定位并抽取出商品的“属性—值”关系。但该方法依赖人工构建的本体库，特别是商品属性本体，无法对本体之外的其他属性进行识别，同时也无法有效利用表格之外的文本信息，造成关系抽取的召回率过低。

文献[2]解决了如何提取、组织网页中热门商品的属性描述文本。首先，基于网页的 DOM 树结构将描述商品属性的文本片段提取出来。然后利用网页的被点击率和同类商品在不同网页间的关联信息构建图模型，最后基于 CRF 模型实现热门商品属性的抽取。该方法仅获得了网页中商品属性的描述文本，并未实现商品属性与值的独立抽取及关联处理。此外，CRF 模型的使用依赖大量人工标注的语料资源，影响实用性和通用性。

文献[4]探讨了如何从商品评论中抽取商品属性，并针对特征的重要性进行排序，借以辅助针对商品评论的情感分析。其中特征抽取部分借助句法分析器提取名词短语作为候选商品属性，然后利用有监督的 SVM 分类器及聚类算法来实现商品特征词的判定。文献[5]通过挖掘评论中的名词性商品属性辅助观点挖掘。文献[6]则提出一种半监督学习算法，用于识别和聚类出商品评论中同一商品属性的不同表述形式。文献[7]提出了基于层级马尔科夫模型的商品命名实体识别方法，可识别出自由文本中的商品命名实体，例如型号、品牌等。但这类实体仅是商品属性的一部分，且此方法不能实现商品的“属性—值”关系识别。

从电子商务网站抽取商品“属性—值”关系也是基于互联网进行特定信息的抽取。构建模板是信息抽取中常用的方法之一（也是本文的基础方法）。模板的构建方法分为两类：有监督和无监督的方法。有监督的模板构建算法多是利用基于人工标注语料的机器学习方法自动构建模板[8]，例如，文献[9-10]利用标记过的网页建立模板。无监督的模板构建算法[11-15]无需人参与，使用网页中重复的模式或者重复的 HTML 标签构建模板。文献[16]通过计算网页信息实例间的相似度，进行单链聚类，通过合并同类别的模板构建出新的模板。但是这一方法依赖于事先确定的信息实例集，并不是完全的无监督方法。文献[17]提出网页链接分类算法和网页结构分离算法，提取出网页中的信息单元，在此基础上构建出各模板。这一方法假设单个网站内的模板不会超过两个，但是在 C2C 电子商务网站中，商品信息的组织方式多样，针对某一网站仅仅学习单个或有限的模板并不能满足实际的抽取要求。因此，本文提出了基于网页标题，并针对每个网页独立学习模板方法。

3　基于模板的“属性—值”关系抽取方法

为了实现商品“属性—值”关系的自动抽取，本文提出了一种结合网页结构信息和字符信息的商品“属性—值”关系模板的自动构建方法。图 2 给出本文方法的工作流程，主要包含如下 4 个关键环节：1）领域属性词包的构建；2）文本节点的细化；3）种子“属性—值”关系的提取；4）优质模板的学习。

为实现优质商品“属性—值”关系的提取，首先需要获得属性词集合。利用网页标题包含属性信息的特点，将标题分词，然后进行汇总。本文将此类词视为潜在商品属性词，忽略词汇的位置关系后构成词集。这一词集被定义为属性词包。

网页文本节点的细化是通过预设的分隔符，并在网页源码中插入特定的字符串进行的。实现了商品属性名与值处于相同和不同文本节点的解决方法的统一。

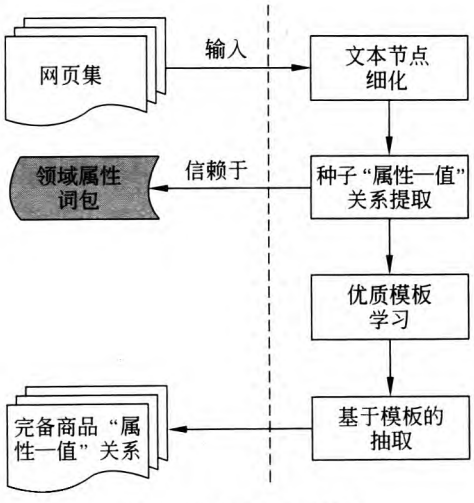

图2 本文方法流程图

在检测高质量“属性—值”关系的过程中，属性词包将用于定位网页中潜在的描述商品属性的文本片段。片段的长度及其中属性词的出现频率用作识别优质片段的特征，筛选出的文本片段经扩展挖掘出种子“属性—值”关系。

优质模板学习部分将结合网页结构信息和字符信息构建模板。一个加权词表被引入对特定的“属性—值”关系及其对应模板进行加权，同时利用HTML标签的对称性对潜在的“属性—值”关系进行了过滤。最后本文提出一个基于模板频率的公式用于学习优质模板。

3.1 领域属性词包的构建

电子商务网站为了提高商品被检索到的概率，倾向于以关键词累加的方式组织网页标题。网页的标题常常包含有商品属性词，如图1中示例网页标题“『原装正品』佳能 IXUS130 数码相机 1 400 万 4 倍变焦 2.7 寸屏”，包含有“佳能”、“数码相机”、“1 400 万”、“4 倍”、“变焦”、“2.7 寸”和“屏”等标识“属性—值”关系的关键词。同时，这类特征词往往高频出现于正文的全文或其局部文本块内。那么，预先从商品标题中抽取足够多的商品属性关键词，然后利用商品属性词从网页内容中检索并定位对应的商品属性文本节点，这类节点有助于精确判定属性词和属性值。然而，单个网页标题的商品属性词数量过少，并且标题中属性描述与网页正文中的表述往往不同，这两类问题将负面影响定位属性片段的准确性。本文通过利用大量网页标题构建商品的属性词包以解决上述问题。

本文从C2C网站（如淘宝、拍拍等）中抽取大量标题。利用中国科学院中文分词工具ICTCLAS对标题进行分词。分词的结果并不进行进一步的处理，直接构成一个词包。在这些标题上的分词效果并不理想，因此虽然利用大量网页标题可以弥补商品属性词稀疏的问题，但也将引入大量的噪音词汇，如示例网页中的“原装”“正品”，以及被错分的“变”“焦”等。这些噪音词会定位出错误的文本节点，在3.3.1节和3.4.2节中将通过基于特征的过滤算法来消除这些错误文本节点的影响。而且通过过滤，本文实验取得极为优异的结果，证明了本文方法的鲁棒性。

3.2 网页文本节点的细分

通过观察解析网页后得到的DOM树结构，我们发现基于属性名与属性值所在的文本节点在DOM树中的不同位置关系类型，商品“属性—值”关系存在两种组织形式：

1）商品属性名和值出现在前后相邻的文本节点中。如图1示例网页中“光学变焦”—“4 倍”在网页HTML中为“＜tr＞＜td＞＜span＞光学变焦＜/span＞＜/td＞＜td＞4 倍＜/td＞＜/tr＞”。

2）商品属性名和值出现在同一文本节点中，如上述示例有可能组织成“＜br＞＜span＞＜span＞光学变焦 ：4 倍＜/span＞＜/span＞”。

已有的从结构化网页中抽取商品“属性—值”关系的方法将DOM解析获得的叶子节点作为最小处理单元，并不能处理第二种情况下的网页。

统计语料得知，在C2C电子商务平台上，商品“属性—值”关系被组织到同一文本节点中的网页占据了较大比例，如果对这类网页不能正确处理，将极大影响最终抽取的效果。进一步分析发现，当卖家将商品的属性名和值组织到同一文本节点时，文本片段的中间通常会存在明显的分隔符，例如，“：”。而在商品属性和值属于不同的文本节点的网页中，这样的分隔符几乎不会出现。因此本文提出对所有文本节点做进一步细化的预处理方案。

本文预定义的分隔符为{“ ”,“：”,“-”}。具体的细分方法如下：对网页中所有的文本节点进行判断，若存在此类分隔符，则将文本节点细分为两段，同时文本节点中的分隔符被特殊字符“＃segment＃”替换。在其后的处理环节中，包含此特殊字符的文本节点将被看作两个独立的文本片段。如HTML源码片段“＜br＞＜span＞光学变焦：4 倍＜/span＞”，将被细分为“＜br＞＜span＞光学变焦

＃segment＃ 4 倍＜/span＞”。

本文定义对所有网页进行细分操作后得到的文本块为文本片段。而细分的作用将在 3.4.1 节中讨论候选模板构建时做详细介绍。

3.3 种子“属性—值”关系提取

本节介绍了如何利用属性词包从细化后的网页中挖掘并筛选出潜在的商品“属性—值”关系。这一模块对应于传统方法中的人工标注语料的过程，通过充分利用现有的资源，为实现完全自动的商品“属性—值”关系挖掘提供基础。

3.3.1 优质属性文本节点的筛选

为了获得优质属性文本片段，所有包含属性词包中词的文本片段将被定位和输出。优质属性文本片段被定义为能精确描述某个属性，且包含较少干扰信息的文本片段。本文利用如下三个特征筛选优质属性文本节点：

1）候选文本片段中包含不同属性词的数目

2）候选文本片段中属性词在整个网页内容中出现的频度

3）候选文本片段的长度

上述特征选择源自如下假设：1）如果一个文本片段包含多个属性词，其描述单个属性信息的能力将会减弱，实验部分检验特定文本片段中包含不同属性词的数量是否超过阈值 α（经验值为 3），如果高于 α 则摒弃这一候选片段；2）相比于常用词汇，商品属性词在正文中出现的频率较低，如果某一特征词在网页中出现的频率很高，其所处的文本片段包含商品“属性—值”关系的概率相应变小，为此，本文设定一个关于特征词出现频率的阈值 β（经验值为 2），频率高于 β 则摒弃这一文本片段；3）如果文本片段过长，其中包含的噪音信息较多，直接影响“属性—值”关系抽取的精确性，因此本文设置阈值 γ（经验值为 6），长度超过这一阈值的文本片段将被弃用。

3.3.2 双向扩展构建种子“属性—值”关系

在商品描述网页中，商品的属性和值总是相邻呈现，如图 3 中所示的“感光元件”和“CCD 传感器”构成“属性—值”关系。因此，必须将获得的优质属性文本片段与相邻文本片段进行组合，借以确定潜在的属性词和属性值搭配关系。其中，由于文本片段的描述对象（“属性”或“值”）难以直接确定，本文采用一种双向扩展方法从优质文本片段中抽取潜在候选关系，并将其用于最终的关系模板构建。

```
<tr>
<td><a target="_blank"><span style="color:#000099,">感光元件</span></a></td>
<td>CCD传感器</td>
</tr>
<tr>
<td><a target="_blank"><span style="color:#000099,">传感器尺寸</span></a></td>
<td>1/2.3英寸CCD</td>
</tr>
<tr>
<td><a target=" blank"><span style="color:#000099 ">总像素数</span></a></td>
```

图 3　示例网页的部分 HTML 源码

双向扩展具体方法如下：在网页源码中每一个优质属性文本片段将分别在前后两个方向取得相邻的文本片段，即每一个将获得两个关系。如图 3 所示，若“CCD 传感器”是一个优质特征文本片段，其前面相邻的文本片段为“感光元件”，后面相邻的文本片段为“传感器尺寸”。它们将分别在一起构成两组关系：“感光元件—CCD 传感器”；“CCD 传感器—传感器尺寸”。这类关系将作为下一模块的输入，下文将其称为种子“属性—值”关系。

3.4 优质模板的学习

模板选择过程将网页中商品“属性—值”关系的提取问题转化为模板构建问题，利用优质关系自动塑造模板。通过对网页文本节点的细化实现了不同情况下模板构建过程的统一。此外，由于双向扩展获得大量噪声关系，因此在模板选择过程中进行筛选，此处侧重利用候选模板的发生频率作为最优模板检测。

3.4.1 候选模板构建与筛选

在商品“属性—值”关系属于不同文本片段的网页中，每一个种子“属性—值”关系在网页中的结构信息对应于 HTML 源码中的标签信息。每一个种子“属性—值”关系在 HTML 源码中都具有三个相邻的 HTML 标签：属性关系之前、之间和之后，模板构建过程选择前、中、后三项标签建立模板。以图 3所示的 HTML 源码段为例，假设“感光元件”和“CCD 传感器”是一项候选属性关系。而在“感光元件”之前是标签串为“＜tr＞＜td＞＜a target＝"_blank"＞＜span style＝"color：＃000099；"＞”，两个文本片段之间的标签为“＜/span＞＜/a＞

</td><td>”，而在“CCD传感器”之后的标签为“</td></tr>”。由此，借助候选属性关系“感光元件”和“CCD传感器”可以构建如下模板：

<tr><td><a>[tf]</td><td>[tf]</td></tr>

其中，“[tf]”用于泛指模板中的“属性词”(前[tf])和“属性值”(后[tf])搭配，如上例中的“感光元件”和“CCD传感器”。由此生成的模板可直接参与后续抽取过程中的HTLM标签匹配，如果匹配成功，两项标签[tf]指代的特征词将被判定为一对“属性—值”关系。

而如果某一被细分成的属性文本片段被扩展为候选属性“属性—值”关系时，其模板构建过程同样由属性关系之前、之间和之后三部分构成，这时被替换的特殊字符串将被视为HTML标签的等价元素。这样上文提到的利用标签的模板构建方法将被直接利用。以细分后的HTML源码片段“
光学变焦 #segment# 4倍”为例，本文方法将构建出如下模板：

[tf] #segment# [tf]

其中“[tf]”的含义同上。

由于对网页源码的细分并插入特殊标记字符串，使得在商品“属性—值”属于不同文本节点和相同文本节点的两种情况下，构建提取商品“属性—值”关系模板的过程得到了统一。网页布局信息和字符信息在本文的模板构建过程中得到融合。

然而双向扩展获得的候选关系往往含有大量噪声。这些噪音信息体现为商品的属性名与前一个属性值，或者某一个属性值与后一个属性名被纳入种子“属性—值”关系中，须将此类候选关系过滤以提高模板构建的准确性，本文方法利用对应模板中的HTML标签的对称性进行过滤。

HTML规范要求大部分标签在使用时必须由开始标签和结束标签配对组成(即对称性特征)。因此，从候选关系中提取的模板标签也需满足这一对称性。实验证明这一特征可有效过滤错误候选关系及其对应的候选模板。

3.4.2 加权词表的引入

某些文本片段的主体部分是由数字和单位词构成，其用于表述商品的某一属性值的概率较大。如在图3中，“1/2.3英寸CCD”就是用来表述“传感器尺寸”这一商品属性的值，其中“1/2.3英寸”构成了这个文本片段的主体。因此为了进一步确保能找到最优模板，本文引入了一个加权词表对可能的优质“属性—值”关系进行加权。如表1所示，选择单位词长度、重量和时间这三种单位词将被作为主题中种子“属性—值”的抽取对象。

表1 加权词表

类型	单位词
长度	m / 米 cm / 厘米 mm / 毫米
重量	kg / 千克 g / 克 mg / 毫克
时间	s / 秒 min / 分 h / 小时 year / 年

加权方法如下：在候选属性关系中，若后一个文本片段的大部分字符是数字以及出现在加权词表中的任一单位词，则对这个属性关系进行加权，其对应模板的频率将会被乘以一个加权系数，在本文中这一加权系数为设定为15。判断一个文本片段中数字和单位词是否占据主体是根据这部分字符的长度占整个文本节点长度的比例。如果其比例超过50%，即认为占据了整个文本片段的主体。

3.4.3 优质模板选择

上述加权处理，可获得一组由HTML标签构成的模板。借助每种模板以及各自出现的统计频率，可有效检测最优模板。在电子商务网站中，一个商品销售网页中通常存在多个描述商品属性的文本块，如图1所示。因此不能简单通过取频率最高值的方式获取优质模板。本文提出了通过设定自适应于特定网页环境(环境包括网页中的模板总数和种类)的动态阈值，过滤错误或噪声模板的方法。阈值计算如式(1)：

$$T = \frac{\sum_{i=1}^{n} Ap_i}{n} \tag{1}$$

其中，n 表示当前网页模板的种类，Ap_i 表示第 i 种模板的出现频率。

加权词表的引入使“属性—值”关系对应模板的出现频率得到了加权，这将提升所有模板出现频率的平均值，那些没有被加权的模板的出现频率将很难超过这一平均值。本文方法认为能够被加权的属性文本对，其对应的模板很有可能是优质模板，因此本文使用所有模板的出现频率的平均值作为阈值，

出现频率大于阈值的模板将被选择为最优模板。

4　实验及分析

4.1　实验语料

本文中实验都在数码相机和手机两个领域上进行。语料由三部分组成：用于构建属性词包的网页标题集；取自 C2C 电子商务网站的网页集；来自普通网站和 B2C 电子商务网站的网页集。

为了构建初始的种子属性词包，针对每个领域分别从“拍拍”网提取 5 000 个网页标题。

此外，通过将商品领域名称（如“数码相机”或“手机”）提交给电子商务网站的搜索引擎，将获得一批相关领域的商品网页。实验中用于抽取商品“属性—值”关系的网页分别来自于“拍拍”和“淘宝”，每个领域分别从检索反馈结果中抽取 4 000 个网页（共 8 000），这些网页构成了实验语料中的取自 C2C 电子商务网站的网页集。将取自 C2C 电子商务网站的网页单独作为一个语料集，是因为其中的网页存在着复杂度高，样式多的特点，相比传统网站存在更大的挑战。

本文中，由于种子属性词包利用了 C2C 电子商务网站中的网页标题进行构建，其对于定位网页中的优质候选片段，以及在此基础上的模板构建都起到了关键的先验作用。然而，C2C 网站中的网页主题包含商品属性的这一鲜明特色，以及借助这一特点的上述“属性—值”关系抽取方法，是否适用于普通的电子商务网站，也是实验需要验证的实用性指标。为此我们从普通网站和 B2C 电子商务网站中抽取网页构成了新的网页集。这类网站包括“中关村在线”、“京东”和“新蛋网”。其中，“中关村在线”是国内最大的科技门户之一，“京东”和“新蛋网”是国内著名的 B2C 电子商务网站。实验均在每个网站“数码相机”和“手机”两个领域分别选取 100 个网页，构成了第三部分语料。

4.2　基准方法

本文中使用了两个基准方法。

第一个基准方法为文献[1]中基于表格本体和属性本体，并结合视觉信息提取商品“属性—值”关系的方法。按照文献中的说明，本文人工构建了这两个本体，其中在数码相机领域的本体包含属性 37 个，手机领域本体包含属性 32 个。

第二个基准方法为不对网页文本节点进行细化处理，以 DOM 树解析中的叶节点为最小单元。通过构造仅由 HTML 标签组成的模板去挖掘商品的“属性—值”关系。

4.3　评价方法

实验中的标准结果由人工标注统计得出，采用准确率 P、召回率 R 和 F 值进行方法性能度量。其中，准确率定义为获得正确商品“属性—值”关系的数目与所有抽取出的商品“属性—值”关系数量的比值；召回率定义为获得正确商品“属性—值”关系的数量与网页中包含商品“属性—值”关系的比值；F 值则为 2PR/(P+R)。

4.4　实验结果及分析

4.4.1　语料统计分析

本文实验中先对基于 C2C 电子商务网站构建的网页集进行了分析。这是由于 C2C 类型的电子商务网站相对于传统的电子商务网站具有如下特点：1）商品种类和数量更为丰富；2）大量卖家组织出来的商品“属性—值”关系信息海量化，使得商品属性描述的多样性得到最大化。通过对语料的量化分析，有助于估计复杂度，为方法选择提供依据。

分别随机从数码相机和手机的网页集中抽出 300 个网页，统计其中商品“属性—值”关系属于同一文本节点和不同文本节点的网页数量之比，结果如表 2 所示。其中数码相机领域网页内，属性与值属于同一文本节点的比例达 14.67%，而在手机领域这一比例高达 24.33%，这一现象证明本文提出的将网页文本节点进行细分的必要性。

表 2　基于 C2C 网站网页的网页集中商品属性与值位置关系统计

	属于同一文本节点	所占比例/%
数码相机	44	14.67
手机	73	24.33

同时，为了调查在 C2C 电子商务平台上网页内容组织的差异性，本文统计了“拍拍”网中随机选择的 1 000 个商品网页，其中，网页组织模式的数量和网页数量的关系如图 4 所示。1 000 个网页中共有 62 种不同商品“属性—值”关系的组织模式，基本呈线性增益关系。这种情况下，若利用机器学习方法自动提取网页商品属性组织模板，其需要的人工标

注语料将很难获得。这一现象表明从此类语料集合上提取商品“属性—值”关系的必要性。

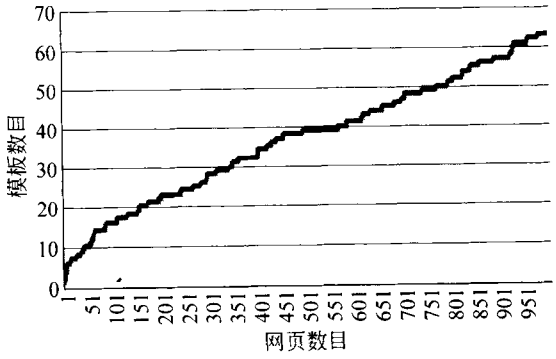

图4 “拍拍”下商品“属性—值”关系组织模式与网页数量的关系

4.4.2 基于取自C2C电子商务网站的网页集的运行结果及分析

在取自C2C电子商务网站的网页集上，每个领域都利用两个基准方法以及本文提出的方法对所有网页进行处理。对于实验结果由人工评测得出，为了尽可能降低人工评测的误差，本文针对上述网页进行10组交叉验证，每组随机选择400个网页进行人工评测。实验结果如表3所示。

表3 基于C2C网站网页的实验结果

	数码相机			手机		
	准确率	召回率	F值	准确率	召回率	F值
基准方法1	0.850 7	0.780 9	0.814 3	0.826 9	0.651 5	0.728 8
基准方法2	0.972 0	0.867 7	0.916 9	0.977 4	0.756 9	0.853 1
本文方法	0.956 7	0.937 0	0.946 7	0.941 6	0.874 4	0.906 8

基准方法1的抽取结果与人工设置的商品属性本体有很大关系，虽然考虑了网页中的视图信息，由于C2C电子商务网站中布局的复杂性，使得此方法在准确率和召回率上均相对较低。基准方法2未对网页的文本节点进行细化，使得商品“属性—值”关系处于同一文本节点中的网页内的商品属性值关系被过滤，导致召回率降低。在基准方法2中，数码相机领域与手机领域的召回率相差达到11.08%，造成这一现象的原因是在相机领域语料中，属性和值被组织到同一个网页中的比例要高很多。相比之下，本文方法在对网页文本节点进行细化后，虽然在两个领域上的准确率有所降低，但是召回率得到了很大的提高。与基准方法2相比，本文方法在数码相机和手机两个领域的F-值分别提高了2.98%和5.37%。

准确率降低的原因主要是一些商品“属性—值”关系组织在不同文本节点中，但是其文本节点中仍含有文本细化中设定的分隔符，造成了文本节点细分过程中的错误。本文中的文本节点细化策略仍较为粗糙，单纯采用了以设定的分隔符进行匹配的方法进行节点细化，对网页内容的变化的适应性不强。如网页源码中存在如下商品信息描述“<td><a>屏幕比例</td><td>16：9</td>”，在本文方法中，“16：9”就被错误切分，进而增加最终商品“属性—值”关系抽取结果中的错误。但是实验依然证明了本文方法中细化文本节点，对于构建融合HMTL标签和字符特征的模板的重要作用。

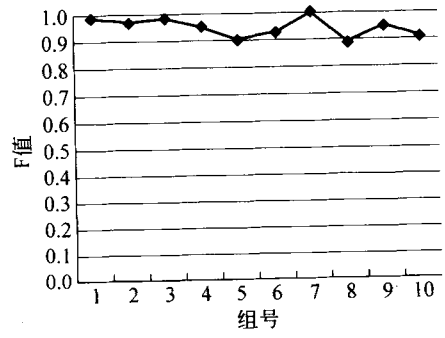

图5 相机评测结果中F值关系图

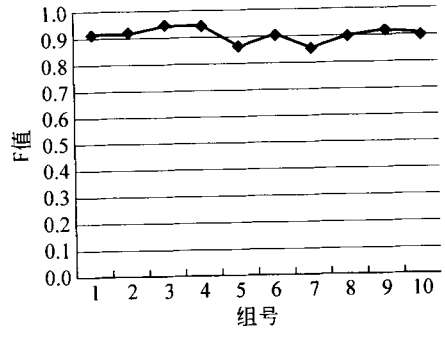

图6 手机评测结果中F值关系图

图5和图6分别显示数码相机和手机领域中，对利用本文方法处理的10组网页F值的变化关系。其横坐标值表示每组评测的组号，纵坐标表示F值。如图所示在“数码相机”和“手机”这两个领域中，每组评测的F值均较接近，这证明了本文方法的稳定性。

4.4.3 基于传统网站和B2C电子商务网站的网页集的运行结果及分析

由于本文方法利用来自C2C电子商务网站的网页标题构建的属性词包，为了验证这一词包的普

适性，本文设计了此组实验。实验均在每个网站“数码相机”和“手机”两个领域分别选取100个网页进行测试。经人工标注和检验，关系抽取性能如表4所示。

表4　基于取自京东，新蛋网和中关村在线的语料的实验结果

	数码相机			手机		
	准确率	召回率	F值	准确率	召回率	F值
京东	0.929	0.999	0.963	0.982	0.999	0.991
新蛋网	0.884	0.998	0.937	0.920	0.999	0.990
中关村在线	0.986	0.999	0.992	0.999	0.927	0.962
平均	0.933	0.999	0.936	0.967	0.975	0.970

结果显示，本文方法在非C2C类的传统网站上取得了更加优异的结果。原因在于传统网站中的网页在组织方式上更一致，相对于C2C电子商务网站，其中模式较少，且干扰的信息也比较少。商品“属性—值”被组织到同一个文本节点中的情况极少发生，因此召回率相对于前一组实验有质的提高。在某些网站，如新蛋网，“属性—值”关系抽取的准确率相对较低，这是由于一些非正文内容的组织样式与商品属性信息相同，而本文方法未区分网页正文与非正文信息，这导致某些错误信息也能被模板匹配成功并抽取出。但这组实验的结果依然能证明本文方法扩展性强，对于一般网页存在较好的泛化能力。

5　总结与展望

本文提出了一种面向网页中商品“属性—值”关系的自动抽取方法。该方法利用对网页中的文本节点进行了细化，并利用大量C2C电子商务网站中的网页标题构建种子属性词包，借助这一词包从网页中定位并筛选出优质属性文本节点。经过双向扩展获得的商品属性关系在筛选和加权后，根据关系的概率分布特征，形成了挖掘和建立融合了页面布局信息和字符信息的模板的统计模型。由此获得的模板在大量网页上（包括C2C和非C2C网站的网页）进行测试，取得了较高的F测度，尤其精确性平均高于0.9。

未来工作将进一步尝试针对非结构化文本的商品“属性—值”关系的自动抽取方法。针对这一问题，本文利用标题定位网页中优质片段并借以生成模板的思想将被继承。但如何从非结构化网页中筛选和提取模板将是有待进一步探索的课题。此外，尝试将本文方法扩展到其他信息领域，并面向更多产品类别进行测试和检验，提高现有方法的健壮性和通用性，也需要进行更深一步的尝试。

参考文献

[1] 赵军，刘康，周光有，等. 开放式文本信息抽取[J]. 中文信息学报，2011，25(6)：98-110.

[2] Wong TL, Lam W. Adapting web information extraction knowledge via mining site-invariant and site-dependent features[J]. ACM Transactions on Internet Technology. 2007,7(1).

[3] Wolfgang Holzinger, Bernhard Krupl, Marcus Herzog. Using Ontologies for Extracting Product Feature from Web Pages[C]//Proceedings of the 10th Annual International Symposium on Wearable Computers (ISWC). 2006: 286-299.

[4] Jianxing Yu, Zheng-Jun Zha, Meng Wang, et al. Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL). 2011: 1496-1505.

[5] Lei Zhang, Bing Liu. Identifying Noun Product Features That Imply Opinions[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL). 2011: 575-580.

[6] Zhongwu Zhai, Bing Liu, Hua Xu, et al. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints[C]//Proceedings of the 23rd International Conference on Computational Linguistics (COLING). 2010.

[7] 刘非凡，赵军，吕碧波，等. 面向商务信息抽取的产品命名实体识别[J]. 中文信息学报，2006，20(1)：7-13.

[8] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization [C]//Proceedings of the 17th National Conference on Artificial Intelligence (AAAI). 2000: 584-589.

[9] Liu L, Pu C, Han W. XWRAP: an XML-enabled wrapper construction system for Web information sources[C]//Proceedings of Internet Conference on Data Engineering(ICDE). 2000: 611-621.

[10] H Elmeleegy, J Madhavan, A Halevy. Harvesting Relational Tables from Lists on the Web[C]//Proceedings of the 35th International Conference on Very Large Data Bases (VLDB). 2010: 1078-1089.

（下转第38页）

1223.

[13] G Gallo,G Longo,S Nguyen. Directed hypergraph and applications[J]. Discrete Applied Mathematics,1993,42:177-201.

[14] 袁毓林.用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法[J].中文信息学报,2005,19(5):37-43.

[15] 陈耀东,王挺,陈火旺.半监督学习和主动学习相结合的浅层语义分析[J].中文信息学报,2008,22(2):70-75.

[16] 刘茂福,李文捷,姬东鸿,等.基于事件项语义图聚类的多文档摘要方法[J].中文信息学报,2010,24(5):77-84.

[17] 王鑫,孙薇薇,穗志方.基于浅层句法分析的中文语义角色标注研究[J].中文信息学报,2011,25(1):116-122.

[18] Chen F, Farahat A, Brants T. Multiple Similarity Measures and Source-pair Information in Story Link Detection [C]//Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Boston, MA, USA: Association for Computational-Linguistics,2004:313-320.

[19] 于江德,李学钰,樊孝忠,等.最大熵模型的事件分类[J].电子科技大学学报,2010,39(4):612-616.

[20] 李荣陆,王建会,陈晓云,等.使用最大熵模型进行中文文本分类[J].计算机研究与发展,2005,42(1):94-101.

(上接第29页)

[11] Arasu A, Garcia-Molina H. Extracting structured data from Web pages[C]//Proceedings of Special Interest Group on Management of Data (SIGMOD). 2003: 337-348.

[12] Zhai Y, Liu B. Web Data Extraction Based on Partial Tree Alignment[C]//Proceedings of the 14th International World Wide Web Conference (WWW). 2005.

[13] Harith Alani, Sanghee Kim, David E. Millard, et al. Automatic Ontology-Based Knowledge Extraction from Web Documents[J]. IEEE Intelligent Systems. 2003. 18(1): 14-21.

[14] Wang J, Lochovsky FH. Data extraction and label assignment for web databases[C]//Proceedings of the 12th International World Wide Web Conference (WWW). 2003: 187-196.

[15] Yang SH, Lin HL, Han YB. Automatic data extraction from template-generated web pages[J]. Journal of Software, 2008, 19(2): 209-223.

[16] 郑家恒,王兴义,李飞.信息模式自动生成方法研究[J].中文信息学报.2004,18(1):48-54.

[17] 梅雪,程学旗,郭岩,等.一种全自动生成网页信息抽取 wrapper 的方法[J].中文信息学报.2008,22(1):22-29.

网页中商品“属性—值”关系的自动抽取方法研究

作者： 唐伟， 洪宇， 冯艳卉， 姚建民， 朱巧明， TANG Wei， HONG Yu， FENG Yanhui， YAO Jianmin， ZHU Qiaoming

作者单位： 苏州大学计算机科学与技术学院,江苏苏州,215006

刊名： 中文信息学报 ISTIC PKU

英文刊名： Journal of Chinese Information Processing

年，卷(期)： 2013,27(1)

参考文献(17条)

1. 赵军;刘康;周光有 开放式文本信息抽取[期刊论文]-中文信息学报 2011(06)

2. Wong TL;Lam W Adapting web information extraction knowledge via mining site-invariant and site-dependent features 2007(01)

3. Wolfgang Holzinger;Bernhard Krupl;Marcus Herzog Using Ontologies for Extracting Product Feature from Web Pages 2006

4. Jianxing Yu;Zheng-Jun Zha;Meng Wang Aspect Ranking:Identifying Important Product Aspects from Online Consumer Reviews 2011

5. Lei Zhang;Bing Liu Identifying Noun Product Features That Imply Opinions 2011

6. Zhongwu Zhai;Bing Liu;Hua Xu Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints 2010

7. 刘非凡;赵军;吕碧波 面向商务信息抽取的产品命名实体识别[期刊论文]-中文信息学报 2006(01)

8. Freitag D;McCallum A Information extraction with HMM structures learned by stochastic optimization 2000

9. Liu L;Pu C;Han W XWRAP:an XML-enabled wrapper construction system for Web information sources 2000

10. H Elmeleegy;J Madhavan;A Halevy Harvesting Relational Tables from Lists on the Web 2010

11. Arasu A;Garcia-Molina H Extracting structureddata from Web pages 2003

12. Zhai Y;Liu B Web Data Extraction Based on Partial Tree Alignment 2005

13. Harith Alani;Sanghee Kim;David E.Millard Automatic Ontology-Based Knowledge Extraction from Web Documents[外文期刊] 2003(01)

14. Wang J;Lochovsky FH Data extraction and label assignment for web databases 2003

15. Yang SH;Lin HL;Han YB Automatic data extraction from template-generated web pages[期刊论文]-Journal of Software 2008(02)

16. 郑家恒;王兴义;李飞 信息模式自动生成方法研究[期刊论文]-中文信息学报 2004(01)

17. 梅雪;程学旗;郭岩 一种全自动生成网页信息抽取wrapper的方法[期刊论文]-中文信息学报 2008(01)