北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.007

面向中文网络百科的属性和属性值抽取

贾真† 杨宇飞 何大可 刘胜久 尹红风

西南交通大学信息科学与技术学院，成都 610031；†E-mail: zjia@home.swjtu.edu.cn

摘要 针对面向中文网络百科条目文章的属性和属性值抽取，提出一种无监督方法。此方法将属性值看作命名实体，利用频繁模式挖掘和关联分析，从文本中抽取类别属性；采用自扩展方法为属性建立触发词表；基于属性触发词和属性值实体标注挖掘属性值抽取模式，利用层次聚类算法获取高质量的模式。在互动百科中采集的数据集上进行实验，结果表明所提方法可行有效。

关键词 知识获取；属性抽取；非结构化文本；模式挖掘

中图分类号 TP391

Attribute and Attribute Value Extracted from Chinese Wiki Encyclopedia

JIA Zhen†, YANG Yufei, HE Dake, LIU Shengjiu, YIN Hongfeng

School of Information and Science Technology, Southwest Jiaotong University, Chengdu 610031;
†E-mail: zjia@home.swjtu.edu.cn

Abstract An unsupervised approach is proposed to extract attribute and attribute value from Chinese wiki encyclopedia entry articles. Attribute values are viewed as named entities and class attributes are extracted based on frequent patterns mining and association analysis. A bootstrapping method is used to find attribute trigger words for each attribute. Attribute value extraction patterns are generated automatically from sentences which contain attribute trigger words and named entity tags of attribute value. Hierarchy clustering algorithm is applied to obtain reliable patterns. Experimental dataset are collected from HudongBaike. The experiment results show that the method is feasible and effective.

Key words knowledge acquisition; attribute extraction; unstructured text; pattern mining

许多互联网应用(如语义搜索、自动问答系统等)都需要知识库作为支撑。依靠专家人工编撰知识库费时、耗力，还存在知识覆盖率低，更新缓慢等诸多问题。如何自动构建大规模知识库是当今研究的热点。类别、属性和实例的属性值是知识库重要组成部分。网络百科(例如维基百科、百度百科和互动百科等)是通过网络用户协作方式创建的大百科全书。网络百科涵盖的类别广泛，条目众多，每个条目名可以看作一个类别的实例，条目文章是对该实例的详细描述，其中包含大量的属性和属性值信息。目前有许多学者基于维基百科研究属性和属性值的自动获取。与维基百科中文版相比，中文网络百科(互动百科和百度百科)的用户数量、条目数量远超过维基百科。据统计，目前互动百科约有 750 万条目，百度百科约有 600 万条目。此外，中文网络百科条目中的信息更加丰富、更新较快，能够反映国内最新的热点事件。因此，面向中文网络百科进行属性和属性值抽取对于中文知识库自动构建具有重要的意义和应用价值。

由于维基百科与中文网络百科的结构不同，现有面向维基百科的知识获取方法对于中文网络百科具有局限性。例如，在维基百科的每个条目中，通常

国家自然科学基金(61170111，61202043，61262058)、中国科学院自动化所复杂系统管理与控制重点实验室开放课题(20110102)和中央高校基本科研业务费专项基金(SWJTU11ZT08)资助

收稿日期：2013–06–20；修回日期：2013–09–13；网络出版时间：2013-11-11 10:25

会有人工标注的、结构化的信息盒。信息盒中包含条目名、属性和属性值。直接从信息盒中抽取属性和属性值是面向维基百科进行知识获取的常用方法[1–2]。为从维基百科条目文章中获取属性值，文献[3–4]利用信息盒，对包含条目名、属性和属性值的句子进行回标，得到属性值抽取的训练语料，为每个属性训练抽取器。然而，中文网络百科仅有少量条目具有信息盒，并且信息盒中往往是同类别条目的通用属性，数量较少，大部分条目特有的属性散落在条目文章中[5]。面向中文网络百科进行属性和属性值抽取的问题亟待解决。

中文网络百科具有分类系统，并且有开放分类标注条目的类别。利用分类系统和开放分类能够将条目进行分类，获得大量的分类文本。本文提出一种无监督的属性抽取方法，首先从给定类别的文本集中抽取属性，为类别建立统一的属性模板，然后挖掘属性值抽取模式，抽取实例的属性值。本文的主要贡献有：1) 提出基于频繁模式挖掘与关联分析的属性抽取方法；2) 提出了基于自扩展的属性触发词抽取方法；3) 提出基于层次聚类的属性模式挖掘方法；4) 以互动百科大学、乡镇、工厂等 6 个类别约 6 万个条目文章为数据源，进行属性和属性值抽取实验，属性抽取的准确率达到 82.23%；属性值抽取综合指标宏平均 F 值达到 83.59%。

对于类别属性的抽取，文献[6]以含有类别词语的大量 Web 文档为数据源，利用词频统计、文本模式和 HTML 标签提取属性词语；文献[7–8]从搜索引擎查询日志中抽取属性；文献[9–12]以种子实例或模式为查询请求，从搜索引擎查询结果页面中抽取属性。对于属性值的抽取，文献[13]根据给定的属性，从半结构化 HTML 文档中抽取实例和属性值；文献[14,15]采用半监督学习，以少量种子产生训练数据，训练属性值抽取器；文献[16]把人物的属性抽取问题转化为实体关系抽取问题，利用支持向量机进行关系判断。文献[17]协作使用条件随机场和支持向量机解决概念实例、属性及属性值的抽取问题。文献[18–22]采用基于知识库的弱监督方法，利用知识库中已有关系实例从未标注数据中产生训练语料。采用有监督或半监督方法进行属性值抽取需要大量的训练语料，基于知识库的弱监督方法依赖于知识库。本文采用无监督方法，利用网络百科数据的海量性和冗余性特点，通过频繁模式挖掘和关联分析获取类别属性；根据属性词语和属性值实体标注从文本中自动获取大量的模式，然后利用模式抽取属性值。

1 属性和属性值抽取方法概述

属性和属性值抽取方法分为以下 5 个阶段，流程如图 1 所示。

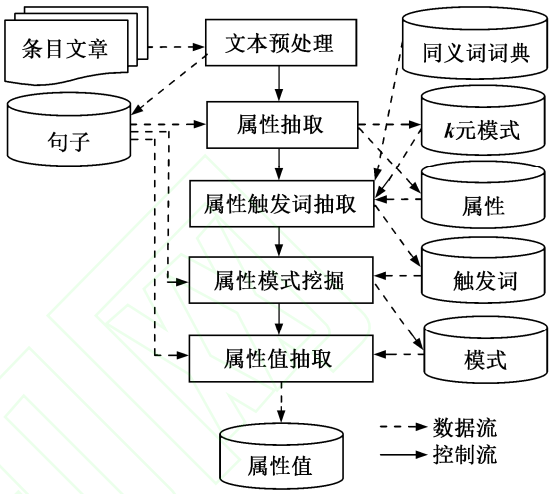

图 1 属性和属性值抽取流程

Fig.1 Pipeline of attribute and attribute value extraction

1) 文本预处理：对百科文章进行分词、词性标注、实体标注和分句等自然语言预处理。

2) 属性抽取：将属性值看作命名实体，从句子中提取 k 元模式，利用 k 元模式分析词语与命名实体标注之间的关联性，与命名实体具有强关联的词语或短语成为属性。

3) 属性触发词抽取：对于每一个属性，利用同义词词典、采用基于自扩展的方法获取属性触发词。

4) 属性模式挖掘：从包含属性触发词和属性值实体标注的句子中挖掘属性值抽取模式，利用层次聚类过滤低频、不可信的模式，获取高质量的模式。

5) 属性值抽取：利用挖掘出的模式从句子中抽取属性值。

2 属性与属性值抽取方法

2.1 文本预处理

本文利用西南交通大学耶宝智慧中文分词平台[23]对条目文章进行了分词、词性标注、实体标注自然语言预处理，并利用汉语分句标点符号对文章进行句子划分。百科条目有一个特点：每个条目页面上的内容是对条目的具体说明，条目文章中常常省略掉被说明的条目，或者用指代词代替条目名。例

如, 大学类别条目文章中常用“该校”或“学校”等指代条目名。我们选择包含条目名或其指代词的段落作为属性抽取语料, 以句号、分号、问号、感叹号和逗号[24]为分句标点符号进行句子划分。以逗号为分句符号时, 需要满足以下两个条件: 划分句子后, 第一, 句子中词的个数不少于 2 个; 第二, 句子中至少含有 1 个名词和 1 个命名实体。不满足以上条件的不能用逗号作为分句符号进行句子划分。

2.2 属性抽取

本文提出基于频繁 k 元模式挖掘与关联分析的类别属性抽取方法。该方法基本思想是: 属性描述语句中往往具有属性词语和属性值, 属性值被看成为命名实体。若某些词语或短语总是和命名实体同时出现, 这些词语或短语与命名实体之间必定具有关系。

2.2.1 k 元模式提取

定义 **1**(句子序列) 经过分词、词性标注与命名实体标注的句子用句子序列表示, 句子序列是由一系列二元组组成的有序序列: $S=<(w_1, t_1), \ldots,(w_i, t_i),\ldots,(w_n, t_n)>$, 其中, w_i 表示词语, t_i 表示 w_i 的词性标注或实体标注, $\forall i \in [1, n]$。

定义 **2**(k 元模式) k 元模式是由 k 个词语、词性标注或实体标注组成的有序序列: $P^k = <x_1, \ldots, x_i, \ldots, x_k>$, 其中, x_i 表示词语、词性标注或实体标注, $\forall i \in [1, k]$。例如<学校,学生>是 2 元模式, <学院,现有,教职工,mq>是 4 元模式, mq 是数量词实体标注。

k 元模式从句子序列中提取。由于属性描述往往是在局部的上下文中, 词语间距离越远, 关联性越小, 我们通过设定窗口限制 k 元模式的提取范围。例如, 句子序列的长度等于 6(即句子中有 6 个词语), 其中 w_3 和 w_5 为命名实体, 窗口的大小为 4, 当窗口中序号为 1, 2, 3, 4 时, 提取的 k 元模式(k=1, 2, 3, 4)有<$\boldsymbol{w}_1$>, <$\boldsymbol{w}_1, w_2$>, <$\boldsymbol{w}_1, t_3$>, <$\boldsymbol{w}_1$, w_4>, <$\boldsymbol{w}_1$, w_2, t_3>, <$\boldsymbol{w}_1$, w_2, w_4>, <$\boldsymbol{w}_1$, t_3, w_4>, <$\boldsymbol{w}_1$, w_2, t_3, w_4>。为避免重复提取, 仅提取窗口中第一个词与其他词组成的模式。w_3 为命名实体, 提取其实体标注 t_3。窗口向后滑动, 当窗口中序号为 2、3、4、5 时, 提取的 k 元模式有<$\boldsymbol{w}_2$>, <$\boldsymbol{w}_2, t_3$>, <$\boldsymbol{w}_2, w_4$>, <$\boldsymbol{w}_2, t_5$>, <$\boldsymbol{w}_2, t_3, w_4$>, <$\boldsymbol{w}_2, t_3, t_5$>, <$\boldsymbol{w}_2, w_4, t_5$>, <$\boldsymbol{w}_2, t_3, w_4, t_5$>。$w_3$ 和 w_5 为命名实体, 提取其实体标注 t_3 和 t_5。窗口逐词向后滑动, 直到句子序列的末尾。

2.2.2 关联分析

定义 **3**(置信度) 置信度用来衡量词语(或短语)与命名实体标注之间的关联程度, 置信度计算公式为

$$\mathrm{conf}(<x_1, x_2, \ldots, x_{k-1}>, <x_k>) = \frac{\mathrm{supp}(<x_1, x_2, \ldots, x_{k-1}>)}{\mathrm{supp}(<x_1, x_2, \ldots, x_{k-1}, x_k>)}, \quad (1)$$

其中$<x_1, x_2, \ldots, x_{k-1}, x_k>(k \geqslant 2)$为 k 元模式, 模式中最后一个元素 x_k 为命名实体标注, k 元模式中的前 $k-1$ 个元素“x_1, x_2, …, x_{k-1}”均为词语, $\mathrm{supp}(<x_1, x_2, \ldots, x_{k-1}>)$为 $k-1$ 元模式的支持度计数, $\mathrm{supp}(<x_1, x_2, \ldots, x_{k-1}, x_k>)$为 k 元模式的支持度计数。

若 $\mathrm{supp}(<x_1, x_2, \ldots, x_{k-1}, x_k>)$大于最小支持度阈值, 且置信度 $\mathrm{conf}(<x_1, x_2, \ldots, x_{k-1}, x_k>, <x_k>)$大于最小置信度阈值, 则模式中的词语“$x_1, x_2, \ldots, x_{k-1}$”组成候选属性词语。例如, <建校, 年代, t>为频繁 3 元模式, 其中, t 为时间实体标注, conf(<建校, 年代>,<t>)大于最小置信度阈值, “建校年代”为候选属性词语, 属性值为时间实体。

2.2.3 属性获取

候选属性中存在部分重复的、冗余的和错误的属性。例如, 表示地理位置的候选属性有“位于”、“地址”和“地处”等等。错误的属性有“固定总价值”、“优秀人数”等, 这些错误有的是由于自然语言预处理引起的, 有的是由于 k 元模式中词语不连续引起的。为了除去错误的和重复的候选属性, 我们采用人工干预的方法, 为每个类别建立统一的类别属性模板。

2.3 属性触发词抽取

定义 **4**(属性触发词) 属性触发词是指可以激活属性值抽取任务的词语。

同一个属性的触发词往往是同义词或近义词, 现有信息抽取系统常用的做法是采用哈尔滨工业大学信息检索研究室的《同义词词林扩展版》[25]扩展触发词。然而, 使用同义词词典扩展触发词存在以下两个问题: 1) 词典中同义词数量有限, 导致触发词的召回率较低; 2) 从词典获取的触发词在文本中并没有出现。为了解决以上两个问题, 本文提出一种基于自扩展的触发词抽取方法, 方法流程如图 2 所示。

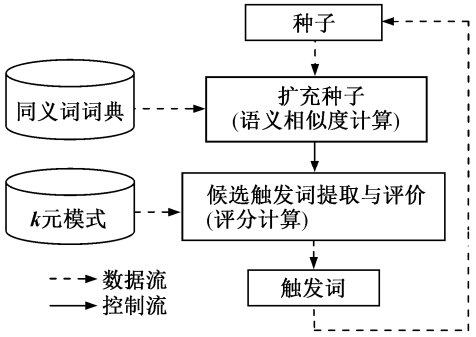

图 2　属性触发词抽取流程

Fig. 2　Pipeline of attribute trigger words extraction

首先，以少量属性触发词为种子，利用语义相似度计算在同义词词林中找出种子的同义词或近义词；然后，从 k 元模式中提取与属性值实体标注共现的词语作为候选属性触发词，计算候选属性触发词的评分，评分高的候选属性触发词成为触发词，新的触发词又作为种子进入下一轮迭代。

2.3.1　扩充种子

对于每一个属性，人工选择若干词语作为种子触发词。由于属性词语可能是组合词，我们采用西南交通大学耶宝智慧中文分词平台对属性词语进行细粒度分词[23]。例如，属性词“学校地址”细粒度分词后为“学校”和“地址”两个词。分别对这两个词语进行同义扩展。例如，“学校”的同义词为该校、母校、全校、学府、院所、院校等，“地址”的同义词为地方、地点等。将每个词语的同义词组合，生成新的种子词，例如“该校地方”、“该校地点”、“母校地方”、“母校地点”等。

本文采用田久乐等[26]的方法进行词语之间的语义相似度计算，满足相似度阈值的词语为同义词。为了减少语义漂移，仅计算第 5 层分支词语间的语义相似度。本文提出的动态生成语义相似度阈值的方法，计算公式如下：

$$\min\_sim(n)=\beta\times\cos\frac{n\times\pi}{180}, \tag{2}$$

语义相似度阈值是分支层节点数目 n 的函数，参数 β 取经验值为 0.95。

2.3.2　候选属性触发词提取与评价

k 元模式$<x_1, \ldots, x_i, \ldots, x_k>$中 x_k 为属性值实体标注，前 $k-1$ 个元素$<x_1, \ldots, x_i, \ldots, x_{k-1}>$中的词为候选属性触发词。为了从候选词语中获取属性触发词，我们提出一种评价方法，给每个候选属性触发词计算一个评分(Score)，评分高的成为属性触发词。Score 值是字面相似度和置信度的加权和，计算公式如下：

$$\begin{aligned}\mathrm{Score(c)}=&\alpha\times\mathrm{Max}\{\mathrm{litSim}(c,s_j),\\ &s_i\in\mathrm{seedSet}\}+\beta\times\mathrm{Conf}(c,p)\end{aligned} \tag{3}$$

式中 α 和 β 为加权值，c 为候选属性触发词，s_i 为扩充种子中的第 i 个种子，Conf 为候选属性触发词序列与属性值实体标注 p 的置信度。经过多次实验，人工评定后将 α 和 β 分别设置为 0.75 和 0.25。

字面相似度计算采用重心后移规律匹配法[27]，计算公式如下：

$$\begin{aligned}\mathrm{litSim}(c,s_i)=&\gamma\times\frac{1}{2}\left(\frac{o}{m}+\frac{o}{n}\right)+\delta\times\min\left(\frac{m}{n},\frac{n}{m}\right)\times\\ &\frac{1}{2}\left(\frac{o}{m},\frac{o}{n}\right)\times\frac{1}{2}\left(\frac{\sum_{k=1}^{o}L_{\mathrm{w}}(k)}{\sum_{t=1}^{m}t}+\frac{\sum_{k=1}^{o}L_{\mathrm{s}}(k)}{\sum_{q=1}^{n}q}\right),\end{aligned} \tag{4}$$

γ 和 δ 为加权值，m 和 n 分别表示 c 和种子 s_i 的字数，o 表示 c 和 s_i 的匹配字数，$L_w(k)$ 和 $L_s(k)$ 分别表示匹配字符 k 在 c 和 s_i 中的匹配序。根据黄金分割律，γ 和 δ 通常定义为 0.6 和 0.4[28]。

2.4　属性模式挖掘与属性值抽取

本文采用模式匹配方法[29]得到实例属性值。包含实例、属性词语和属性值的句子都是潜在的属性值抽取模式。由于百科文本中常省略条目名，为了提高属性值的召回率，本文以属性触发词与属性值实体标注之间的文本作为属性值抽取的候选模式。例如“校长”为属性词语，nr 为属性值实体标注，从句子序列<(严几道,nr),(先生,n), (为,p), (本校,r),(校长,nnt),(时,ng)>抽取候选模式<nr, 先生，为，本校，校长>。

候选模式中有一些是噪声，为了提高属性值的准确率需要对候选模式进行筛选，找出高质量的模式。本文提出一种基于层次聚类的模式筛选方法：1) 将每个属性的所有模式按照属性触发词进行分类；2) 利用层次聚类将分类后的模式进行聚类；3) 聚类后，若某个簇中不同模式的个数少于阈值，并且簇中每个模式的支持度计数小于阈值，则将该簇中的所有模式删去。

层次聚类算法如下。

输入：模式集合 P；相似度阈值 min_sim；簇中最少模式个数 min_count；模式最小支持度 min_supp；

输出: 聚类后得到的簇 Cluster = {cluster$_1$, cluster$_2$, …};

1 初始化簇集合 Cluster = {cluster$_1$, cluster$_2$, …, cluster$_n$}(cluster$_i$ = { p_i }, $p_i \in P$);

2 计算簇间距 Dis(cluster$_i$,cluster$_j$);

3 如果簇间距小于 min_sim, 则将两个簇合并, 返回第 2 步;

4 对于 Cluster 中的每个簇 cluster$_i$, 如果 cluster$_i$ 中的模式个数小于 min_count 并且 cluster$_i$ 中每个模式的支持度小于 min_supp, 则将 cluster$_i$ 删除;

5 返回 Cluster。

簇间距为簇中所有模式两两之间相似度的最大值, 簇间距计算公式为

$$\mathrm{Dis}(c_i, c_j) = \max\{\mathrm{pSim}(a, b), a \in c_i, b \in c_j\}, \quad (5)$$

其中 a, b 分别为簇 cluster$_i$ 和 cluster$_j$ 中的模式, pSim(a, b)为模式相似度。模式相似度计算公式为

$$\mathrm{pSim}(a,b) = 1 - \frac{\mathrm{ed}(a, b)}{\max(\mathrm{length}(a), \mathrm{length}(b))}, \quad (6)$$

ed(a,b)是模式 a 和 b 的编辑距离。

我们利用筛选出的模式从句子中抽取属性值。对于同一个属性触发词的模式, 按照长模式优先顺序进行匹配。一旦任意属性触发词的模式匹配成功, 就抽取一个属性值。

3 实验与结果分析

3.1 实验数据集

我们从互动百科下载了大学、中小学、乡镇、行政村、工厂、公司 6 个类别, 共约 6 万个条目作为实验数据。实验分为两部分: 属性抽取和属性值抽取。

3.2 属性抽取实验

我们对文本中的实体标注进行统计, 并以出现次数最多的 5 类实体[23] (数量词、地名、人名、时间和农作物)作为属性值实体类型, 在给定窗口范围内提取频繁 k 元模式、计算置信度、获取候选属性, 然后人工干预生成类别属性。窗口的大小对属性抽取结果有一定影响, 当窗口较小(如窗口为 3)时, k 元模式的数量较少, 候选属性和生成的属性个数较少; 当窗口逐渐增大时, k 元模式的数量增多, 候选属性和生成的属性个数都会增多。但是窗口越大, 错误的候选属性越多, 造成准确率下降。通过实验发现, 窗口为 5 时获得的属性个数较多、准确率较高。准确率计算公式为

$$\text{准确率}(P) = \frac{\text{正确的属性个数(含重复属性)}}{\text{候选属性个数}} \times 100\%。 \quad (7)$$

当窗口为 5 时, 6 个类别的属性准确率统计以及最终生成的属性示例(按置信度排名前 10 位)如表 1 所示。

互动百科对“学校”、“地区”和“公司”等类别人工定义了统一的属性名称。“大学”和“中小学”、“乡镇”和“行政村”、“工厂”和“公司”分别使用“学校”、“地区”和“公司”3 个类别的属性模板。我们用覆盖率(重合属性个数与人工定义属性个数之比)说明本文方法的性能, 如表 2 所示。

从表 2 看出, 互动百科人工定义的属性数量较少, 并且是一些通用属性, 缺少类别特有的属性, 不能体现类别的特征。本文直接从分类文本中获取

表 1 类别属性抽取准确率与属性示例

Table 1 Accuracy of class attributes extraction and examples of attributes

类别	准确率/%	属性示例
大学	73.67	学校地址, 建校年代, 在校学生, 现有教职工, 纸质图书, 本科学制, 专任教师, 学校占地面积, 校舍建筑面积,校长
中小学	72.43	学校地址, 学校占地面积, 始建于, 现有教职工, 联系人, 现有教学班, 一级教师, 校长, 二级教师,建筑总面积
乡镇	65.46	位于, 总面积, 总人口, 农民人均纯收入, 辖行政村, 总户数, 主产, 非农业人口, 农业总产值, 人均耕地面积
行政村	69.17	位于, 隶属于, 人均耕地, 外出务工, 主要种植, 农户数, 年平均气温, 农田面积, 农村经济总收入, 种植业收入
工厂	81.47	员工人数, 法定代表人, 地址, 成立时间, 年营业额, 厂房面积, 月产量, 年出口额, 注册资本, 联系人
公司	82.23	实现销售收入, 公司总部位于, 董事长, 成立日期, 注册地址, 集团董事局主席, 集团总裁, 员工, 股票代码, 营业收入

表 2 类别属性覆盖率统计

Table2 Coverage of class attributes extraction

类别	学校		地区		公司	
	大学	中小学	乡镇	行政村	工厂	公司
人工定义属性个数	8	8	17	17	17	17
本文抽取属性个数	55	43	83	96	25	23
重合属性个数	5	5	11	10	7	10
覆盖率	62.5%	62.5%	64.7%	58.8%	41.2%	58.8%

属性, 属性个数较多, 更能体现类别的特征。而且本文抽取的属性都是文本中存在的, 为属性值抽取提供了依据。但与人工定义属性相比, 本文方法的覆盖率较低, 原因在于本文方法只能抽取属性值为实体标注的属性, 对于人工定义的某些属性, 例如校训(学校)、地标(地区)、性质(公司)等属性, 则无法抽取。

3.3 属性值抽取实验

我们对 6 个类别的实例进行属性值的抽取。本文采用准确率、召回率和 F 值作为属性值抽取的评价标准, 计算公式如下:

$$准确率(P)=\frac{正确的属性值个数}{抽取属性值个数}\times 100\%, \quad (8)$$

$$召回率(R)=\frac{正确的属性值个数}{属性值个数}\times 100\%, \quad (9)$$

$$F值(F\text{-score})=\frac{2\times P\times R}{P+R}\times 100\%。 \quad (10)$$

评价过程如下: 每个类别随机选取 200 个文本, 对文本中的属性和属性值进行人工标注。分别对每个属性的属性值抽取结果进行评价。受篇幅限制, 本文仅列出大学类别的属性值抽取实验结果, 如表 3 所示。

从表中看出, 在校学生、现有教职工和专任教师的准确率较低, 造成抽取准确率较低的原因主要有 3 个: 1) 触发词扩展引起语义漂移, 某些触发词与属性词并不是同义词, 例如学生和留学生, 教师和教职工等; 2) 由于条目文章中多处出现某属性的描述信息, 例如大学条目文章中对院、系描述的段落里也有教师或学生的属性信息, 从而造成一个条目的某属性有多个属性值; 3) 模式错误, 例如从句子序列<(在校学生, nnd), (来自, v), (20 多个, mq), (国家, n)>中抽取的模式<在校学生,来自,mq>是错误的模式。此外, 分词和实体标注自然语言预处理错误对抽取结果也有较大影响。例如校长属性的属性值是人名, 若人名识别错误, 会造成准确率和召回率下降。

表 3 大学类别属性值抽取实验结果

Table 3 University attribute valuesextractionresults

属性	属性值总个数	准确率/%	召回率/%	F 值/%
学校地址	3979	81.23	80.33	80.78
建校年代	4103	76.26	81.76	78.91
在校学生	13361	**51.45**	90.43	65.59
现有教职工	2657	**53.02**	90.78	66.94
纸质图书	3503	77.35	86.11	81.50
本科学制	3508	74.57	90.02	81.57
专任教师	8005	**61.39**	85.78	71.56
学校占地面积	581	71.96	88.66	79.44
校舍建筑面积	508	66.89	90.75	77.01
校长	1840	76.06	72.13	74.04

我们用宏平均计算 6 个类别的属性值抽取准确率、召回率和 F 值, 属性值抽取性能实验结果如表 4 所示。

从实验结果看, 本文提出的方法能够取得较好的召回率, 然而准确率还有待进一步提高。

由于缺乏人工标注的训练语料, 难以与现有的有监督方法进行比较。文献[3]利用维基百科信息盒, 在句子中对实例、属性和属性值进行回标获取训练语料, 用分类器对训练语料进行优化, 并训练 CRF 抽取模型。我们实现了文献[3]的方法, 但由于该方法依赖于信息盒, 我们在同样的数据集上, 对大学类别信息盒中的“所属地区”(学校地址)、“创建时间”(建校年代)等 5 个属性进行了抽取实验。实验结果如表 5 所示。

从表 5 看出, 文献[3]方法的准确率比本文方法高, 但召回率较低, 总体性能比本文方法差; 并且文献[3]方法受到信息盒的局限, 仅能抽取信息盒中的属性。因此, 本文方法更加适用于面向中文网络

表 4 6 个类别属性值抽取实验结果

Table 4 Six classes attribute valuesextraction result

类别	宏平均准确率/%	宏平均召回率/%	宏平均 F 值/%
大学	69.02	85.68	77.35
中小学	71.23	87.68	79.46
乡镇	69.44	88.76	79.10
行政村	71.76	89.27	80.52
工厂	75.63	91.54	83.59
公司	72.49	90.03	81.26

表 5 属性值抽取结果比较

Table5 Comparison of attribute valuesextraction results

属性	本文方法			文献[3]方法		
	准确率/%	召回率/%	F 值/%	准确率/%	召回率/%	F 值/%
学校地址	81.23	80.33	80.78	87.82	63.62	73.79
建校年代	76.26	81.76	78.91	84.91	69.83	76.64
校长	76.06	72.13	74.04	78.82	40.49	53.50
学生人数	51.45	90.43	65.59	80.81	53.51	64.39
教师人数	61.39	85.78	71.56	82.32	44.77	58.00

百科的属性抽取。

4 结论

本文将属性值看作命名实体, 从中文百科分类文本中提取类别属性, 采用自扩展方法获取属性触发词, 利用属性触发词和属性值命名实体标注自动获取属性值抽取模式, 通过过滤不可信的信息抽取模式提升模式质量, 最后利用模式自动抽取属性值。本文方法不依赖人工标注的训练语料, 具有良好的移植性, 有效解决了面向中文网络百科自由文本的属性和属性值抽取问题。本文下一步将研究如何进一步提高属性值抽取的准确率和召回率, 构建能够面向实际应用的中文网络百科知识获取系统。

致谢 研究工作得到杨燕教授、李天瑞教授以及廖浩伟、左玲、陈方正、吴安峻、柏玉同学的帮助, 表示衷心感谢。

参考文献

[1] Suchanek F, Kasneci G, Weikum G. Yago: a core of semantic knowledge unifying WordNet and Wikipedia //Proc of WWW 2007. New York:ACM,2007: 697–706

[2] Auer S, Bizer C, Lehmann G, et al. DBpedia: anucleus for a Web of open data//Proc of ISWC 2007. Berlin: Springer, 2007: 722–735

[3] Wu Fei, Weld D. Autonomously Semantifying Wikipedia // Proc of CIKM 2007. New York:ACM, 2007: 41–50

[4] Wu Fei, Weld D. Automatically Refining the Wikipedia Infobox Ontology//Proc of WWW 2008. New York: ACM, 2008:635–644

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

[6] Tokunaga K, Kazama J, Torisawa K. Automatic discovery of attribute words from web documents // Proc of IJCNLP 2005. Berlin: Springer, 2005: 106–118

[7] Paşca M. Organizing and searching the world wide web offacts-step two: Harnessing the wisdom of the crowds // Proc of WWW 2007. New York: ACM, 2007: 101–110

[8] Paşca M, Durme B. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs // Proc of ACL 2008. Stroudsburg: ACL, 2008: 19–27

[9] Kopliku A, Sauvagnat K, Boughanem M. Retrieving attributes using web tables//Proc of JCDL 2011.New York:ACM, 2011:13–17

[10] Sanchez D. A methodology to learn ontological attributes from the web. Data and Knowledge Engineering, 2010, 6(69): 57–597

[11] 康为, 穗志方. 基于 Web 弱指导的本体概念实例及属性的同步提取. 中文信息学报, 2010, 24(1): 54–59

[12] 李文杰, 穗志方. 基于并列结构的概念实例和属性的同步提取方法. 中文信息学报, 2012, 26(2): 82–87

[13] Yoshinaga N, Torisawa K. Open-domain attribute-value acquisition from semi-structured texts // Proc of the Workshop on Ontolex 2007. Berlin: Springer, 2007: 55–66

[14] Probst K, Ghani R, Krema M, et al. Semi-supervised learning of attribute-value pairs from product descriptions // Proc of IJCAI2007. Berlin: Springer, 2007: 2838–2843

[15] Bakalov A, Fuxman A, Talukdar P,et al.Scad: collective discovery of attribute values // Proc of WWW 2011. New York: ACM, 2011: 447–456

[16] 郭剑毅, 李真, 余正涛, 等. 领域本体概念实例、属性和属性值的抽取及关系预测. 南京大学学报: 自然科学版, 2012, 48(4): 383–389

[17] 叶正, 林鸿飞, 苏绥, 等. 基于支持向量机的人物属性抽取. 计算机研究与发展, 2007, 44(增刊): 271–275

[18] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data // Proc of ACL 2009. Stroudsburg, PA: ACL, 2009: 1003–1011

[19] Yao Limin, Riedel S, McCallum A. Collective cross-document relation extraction without labeled data // Proc of EMNLP 2010. Stroudsburg, PA: ACL, 2010: 1013–1023

[20] Riedel S, Yao Limin, McCallum A. Modeling relations and their mentions without labeled text. Machine Learning and Knowledge Discovery in Databases, 2010, 6323: 148–163

[21] Surdeanu M, McClosky D, Tibshirani J, et al. A simple distant supervision approach for the TAC-KBP slot filling task // Proc of the TAC-KBP 2010 Workshop. Gaithersburg, MD, 2010

[22] Hoffmann R, Zhang C, Ling Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations // Proc of ACL-HLT 2011. Stroudsburg, PA: ACL, 2011: 541–550

[23] 尹红风,贾真,李天瑞. 西南交通大学耶宝智慧中文分词平台[OL]. (2012–07) [2012–11]. http://www.yebol.com.cn

[24] 李艳翠, 冯文贺, 周国栋, 等. 基于逗号的汉语子句识别研究. 北京大学学报: 自然科学版, 2013, 49(1): 7–14

[25] Che Wanxiang, Li Zhenghua, Liu Ting. LTP: A Chinese language technology platform//Proc of the Coling 2010: Demonstrations. 2010: 13–16

[26] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法. 吉林大学学报: 信息科学版, 2010, 28(6): 602–608

[27] 张雪英, 闾国年. 基于字面相似度的地理信息分类体系自动转换方法. 遥感学报, 2008, 12(3): 433–440

[28] 王源, 吴晓滨, 涂从文, 等. 后控规范的计算机处理. 现代图书情报技术, 1993(2): 4–7

[29] Hearst M. Automatic Acquisition of Hyponyms from Large Texts Corpora//Proc of COLING 1992. Stroudsburg: ACL, 1992: 539–545