基于规则的百科人物属性抽取

李红亮 杨 燕 尹红风 贾 真

（西南交通大学信息科学与技术学院 成都 610031）

摘 要 信息抽取是数据挖掘的一个重要领域，文本信息抽取是指从一段自由文本中抽取出指定的信息并将其结构化数据存入知识库供用户查询或下一步处理所用。人物属性信息抽取是智能人物类搜索引擎构建的重要基础，同时结构化信息也是计算机所能理解的一种数据格式。作者提出了一种自动获取百科人物属性的方法，该方法利用各属性值的词性信息来定位到百科自由文本中，通过统计的方法发现规则，再根据规则匹配从百科文本中获取人物属性信息。实验表明该方法从百科文本中抽取人物属性信息是有效的。抽取的结果可以用来构建人物属性知识库。

关键词 人物属性抽取；规则获取；自由文本

Rules-Based Character Attributes Extraction from Baidu Encyclopedia

LI Hong-liang YANG Yan YIN Hong-feng JIA Zhen

(*School of Information Science & Technology, Institute of Noetics and Wisdom, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract Information extraction is an important area of data mining. Text information extraction means extracting specified information from a section of free text and storing structured data in the knowledge base for user querying or further processing. Character attribute information extraction is an important instrument of building search engine of persons, and is also a technology for computer program understanding. This paper presents an automatic method to obtain encyclopedia character attributes, and this method uses the speech tagging of each attribute value to locate the encyclopedia free text. The rules are discovered by statistical method, and the character attributes information is obtained from encyclopedia text according to rules matching. Experiments show that this method is effective in extracting character attribute information from encyclopedia text. The extracted results can be used to build the knowledge base of the character attributes.

Keywords character attributes extraction; rules acquisition; free text

1 引 言

目前人们生活在一个网络信息的时代，互联网已经融入到了人们生活的每一个角落。面对互联网信息的繁杂、无序，人们急需一种有效快捷的信息获取方式。同样，在人物搜索方面要想做到快捷、有效，就需要构建相关知识库，而人物基本属性抽取是人物知识库构建的一个重要基础技术，同时也是人物领域智能搜索方面的一个重要基石。有规则的数据才能被计算机所理解和应用，因此从自由文本中抽取人物属性信息已经成为一个重要的课题。如何从自然表达的文本信息中抽取出这些结构化的人物属性信息是本文研究的主要目的，本文的工作关注于人物属性的抽取，其研究成果可以服务于人物信息检索、智能搜索等方向。

基金项目：国家自然科学基金(61152001，61170111)；中国科学院自动化研究所复杂系统管理与控制重点实验室开放课题(20110102)；中央高校基本科研业务费专项资金(SWJTU11ZT08)。

作者简介：李红亮，硕士研究生，主要研究方向为数据挖掘，E-mail：lihongliangmy@126.com；杨燕，博士，教授，主要研究方向为数据挖掘、计算智能、集成学习等；尹红风，博士，教授，主要研究方向为大数据处理、语义网、搜索引擎；贾真，硕士，讲师，主要研究方向为数据挖掘、智能信息处理。

一个人的通用属性信息一般包括一个人的姓名、性别、出生日期、籍贯、出生地、逝世日期、逝世地点、毕业院校、现任职务、履历信息等内容。目前信息抽取的主要方式有两种：一种是基于规则的方式，第二种是基于统计的方式。基于规则的信息抽取是一个学习和应用的两阶段过程，包括规则的学习和应用规则获取目标信息。信息的抽取规则主要来源于目标上下文约束环境，只要在文本中找到满足规则的约束信息，也就达到了信息抽取的目的，因此规则本身的学习和提取成为基于规则信息抽取的关键。基于统计的信息抽取准确性一般较低，但是有较好的移植性，一般对领域知识的要求不高。比较常见的统计模型有：隐马尔可夫模型、条件随机场、概率上下文无关等，目前已具备较强的统计理论基础，已有健全的训练算法，但是基于统计的信息抽取需要标注好的训练数据。

目前在中文类文本中做信息抽取的还比较少，还没有较成熟的技术。叶正、林鸿飞[1]等把人物属性抽取作为实体关系抽取的一种具体应用，通过 Hownet 获取描述人物属性的触发词，将触发词和人名间的描述关系转化为分类问题，该方法在训练分类器时需要人工标注数据，同时还需用到语义资源。王全剑[2]等在基于句子分块和命名实体识别标记的抽取模式基础上，利用 Wikipedia 作为知识库，提出基于当前元组与关系表示集合语义相似度的关系判别算法对按照模式抽取得到的关系元组进行过滤和分类。丁君军[3]等通过人工制定规则来实现学术概念属性的抽取。车万翔[4]等使用基于特征向量的机器学习算法 Winnow 和支持向量机(SVM)的算法对实体关系进行抽取。陆科新[5]等提出通过构建本体的方法实现IT业招聘应聘这一领域的信息抽取。基于规则[6,7]的抽取方法需要选取构成规则的关键特征词，特征词[8]的选取对规则的生成有着直接的影响。鉴于百度百科人物文本描写语言的集中性并且缺少基本的标注语料，本文采取基于规则的人物属性获取方法。

2 百度百科人物属性获取方法

人物属性的描述复杂多变、形式不一、长短各异，中间可能还包括形容词、副词、助词等词语，这些特点使得在人物属性抽取方向有一定的困难。另一方面语言的语法特性和书写习惯也使得在人物属性描述方面有一定的内在规律。属性值前 n 个词和后 m 个词的描述会存在一定的规律，因此作者通过统计的方法来发现属性值前后的规律，再通过进一步的精化、合并等计算制定最后的抽取规则。最后再将过滤掉的规则集通过频繁项集的计算进行合并选取频率高的规则集作为后备抽取规则。

2.1 规则的获取与算法描述

从自由文本中自动抽取人物属性的关键在于如何自动构造描述人物特定属性的规则，候选规则集的生成主要依赖于人物属性值的词性标记以及生成规则时左右词的特征，根据属性值词性标记定位到待抽取文本中含有该词性标记的词。生成规则时，主要选取该词性标记前后的词为特征，统计每一个候选规则出现的频率。

候选规则集生成算法描述如下：

输入：属性值词性标记/x，待抽取文本 C

输出：候选规则集 G

(1)将待抽取文本读入程序并分词；

(2)用属性值词性标记/x 在 C 中去匹配；

(3)在/x 的前面和后面分别取 m 和 n 个词，生成规则；

(4)如果该词性标记后面的词性标记与该属性标记相同则不计算该词个数；

(5)如果往前查找碰到“，”、“。”、“；”、“！”等终断标点时则停止搜索，同时保留该标点作为特征；

(6)若/x 后面的特征词出现“
”或“，”或“！”或“；”或“？”或“。”等终断标点时，该标点和标点后面的词不再保留；

(7)若规则中只包含/x、连词、介词、数量词、标点、形容词、副词、助词、方位词、代词等无意义的词时，则删除该规则(具体根据不同的属性)；

(8)根据规则出现频率排序规则。

本算法中的分词工具使用西南交通大学思维与智慧研究所开发的中文分词软件， 其网址为：http://www.yebol.com.cn。

2.2 规则的精化

2.2.1 关键词的选取

关键特征词的选择就是发现重要的语境词，这些关键特征词可以区分关系抽取的类型，过滤掉无意义的候选规则。经过分析候选规则发现，每一特定属性规则都有一个核心的关键词或触发词。比如，在出生地属性中可能含有关键特征词“出生于”、“生于”等。基于以上分析，本论文中使用基于熵

的特征选取方法来发现关键特征词。其算法可以如下描述：令 $G=\{g_1,g_2,g_3,\cdots,g_n\}$为所有候选规则的集合，$W=\{w_1,w_2,w_3,\cdots,w_n\}$是所有候选规则集中词的集合。

假定 G_n，$1\leqslant n\leqslant N$，W 的特征空间为 M，则 G 的熵值的计算公式可以概括如下：

$$E=-\sum_{i=1}^{N}\sum_{j=1}^{N}(S_{i,j}\log^{S_{i,j}}+(1-S_{i,j})\log^{(1-S_{i,j})}) \quad (1)$$

(1)其中 $S_{i,j}$ 表示 G_i，G_j 之间的相似度，G_i，G_j 之间的距离用欧氏距离表示，该算法从特征词空间中依次删除每一个词，然后在新的特征词空间中通过公式 1 计算数据的熵E，通过计算可以得到熵的集合$\{E_1,E_2,\cdots,E_m\}$，再对它们排序，E_i 值越大则词 w_i 越重要。基于以上规律我们选取前 K 个词作为规则的关键特征词。

2.2.2 规则的过滤与分类

通过选择的关键特征词集合，可以将不含有关键特征词的候选规则删除，减少候选规则的数量并提升下一步的计算效率，由于第一步获取的候选规则集中可能包含多个属性的规则，所以要根据关键特征词代表的属性类别，将规则集分类。具体算法步骤可描述如下：

输入：规则集 $G=\{g_1,g_2,g_3,\cdots,g_n\}$，关键特征词集合 $K=\{k_1,k_2,k_3,\cdots,k_m\}$

输出：规则集的分类 $Class=\{c_1,c_2,\cdots,c_m\}$

```
Begin
    For i∈[1,m]do
        For p∈P do
            If p 包含 ki
                then ci=ci∪p
    return Class
End
```

2.2.3 规则的生成

由上述算法产生的规则集 G 中还存在一些无意义的规则排在前面的情况，和一些有意义的规则排在后面的情况。为了解决这些问题，需要从中选择一部分质量好的规则，生成最终的规则集作为某一个属性的规则集，具体的算法可用如下公式说明：

$$\text{support}(X_{l1}X_{l2}\cdots X_{ln}X_{r1}X_{r2}\cdots X_{rn}\rightarrow Y)=\frac{P(X_{l1}X_{l2}\cdots X_{ln}X_{r1}X_{r2}\cdots X_{rn})}{P(Y)} \quad (2)$$

其中 X_{li} 代表某一规则中属性值标记左边的一个特征词，X_{ri} 代表某一规则中属性值标记右边的一个特征词，Y 代表该规则中的属性值词性标记。将规则按照该结果再次排序，选取排在前面的规则集合作为该属性的规则集。将过滤掉的规则再次根据频繁项集进行合并，选取频率高的规则作为后备抽取规则，当规则集中抽取不到相应的属性时就使用后备规则集进行文本抽取。

2.3　基于规则的人物属性抽取

对待抽取的每个属性，用规则集中的每个规则分别到自由文本中去匹配，匹配时优先选择频率高的规则。这在一定程度上可以提高抽取的准确率同时可以降低规则匹配到无关句子的数目。如果所有规则没有匹配到相应的属性时，则使用后备规则去匹配。以某一个属性的抽取算法为例，描述如下：

输入：待抽取文本，抽取规则集，后备抽取规则集

输出：属性值

(1)读取待抽取文本的一个句子；

(2)用规则集中的第 i 个规则来匹配该句子(规则集按频率降序存储，i 初值等于 0)；

(3)如果匹配成功，则保存该结果，退出程序；

(4)如果匹配不成功，则 i+1，如果还有没被匹配的规则，转向 2，否则转向 5；

(5)用候选规则集匹配；

(6)如果匹配成功，则保存该结果，退出程序；

(7)如果匹配不成功，并且该待抽取文本还没有结束，则转向 1，否则退出程序。

3　实验结果与分析

3.1　实验数据

本文实验中的数据采用百度百科 2012 年 2 月 10 日的网页数据文件，该数据记录了所有在线百度百科的人物类文本信息。

实验过程中使用到的数据是文本格式数据，所以需对下载的网页数据进行网页信息抽取转换成文本格式。实验中使用的分词、词性标注系统由西南交通大学思维与智慧研究所开发。

3.2　实验评测与结果

本次实验的语料来自百度百科的人物信息类文本，采用准确率(P)和查全率(R)两项指标来衡量算法的性能。

$$P=\frac{C_1}{C_2}\times 100\% \quad (3)$$

$$P=\frac{C_1}{C_3}\times 100\% \quad (4)$$

其中，C_1 代表抽取出来的某一属性的正确的属性值的个数，C_2 代表抽取出来的某一属性值的个数，C_3 表示语料中含有的某一属性值的总数。评价过程中，对任意属性关系，我们随机抽样大小为 300 和 500 的实例结果，统计其准确率和查全率。通过结果分析来说明算法的稳定性和有效性。实例为 300 时的统计结果如下表 1 所示。实例为 500 时的统计结果如下表 2 所示。

表 1 人物属性抽取结果

属性	性别	出生地	出生日期	籍贯	逝世地点	逝世日期	毕业院校	毕业时间
准确率	86.4%	83.6	82.3%	76.2%	85.7%	81.8%	82.8%	77.6%
查全率	84.2%	82.5%	80.8%	77.3%	83.5%	82.4%	79.7%	78.9%

表 2 人物属性抽取结果

属性	性别	出生地	出生日期	籍贯	逝世地点	逝世日期	毕业院校	毕业时间
准确率	86.2%	84.1	82.1%	76.4%	85.5%	81.4%	82.2%	77.2%
查全率	83.6%	82.1%	81.6%	75.3%	82.6%	81.5%	79.2%	76.8%

从以上结果可以看出，籍贯和毕业时间的准确率和查全率较低，主要是由于这两类属性的表达方式多样，直接导致抽取规则生成的限制。其他属性的表达方式较为固定，规则的生成也相对较容易。通过表 1 和表 2 结果的对比和分析，证明该算法的稳定性很好，通过实验结果的统计证明该算法。

通过分析结果得知，影响准确率的主要原因有两点：一是在句子抽取过程中没有考虑指代消解的问题；二是分词软件的标注错误问题。影响查全率的主要原因是信息的抽取规则还不够完整。其中错误的分词标注对两者都有影响。由于中文自然语言书写和表示的丰富和灵活性，直接给信息抽取带来了较大的复杂性和巨大的困难性，所以这方面的研究还需更进一步的突破和努力。

4 结束语

本文提出了一种抽取百科人物属性的方法。将属性值的词性标注信息标记到百度百科文本中，通过统计该标记前后词的出现频率和进一步的计算来生成抽取规则，从而实现人物属性信息的自动抽取，该方法利用了待抽取的属性值的词性标记来获取抽取规则，最后，在百度百科的人物类数据集上进行了实验。实验结果表明，该方法对于抽取百度百科文本中的人物属性具有较好的效果。

本文未来的工作：(1)利用指代消解技术，进一步精化抽取的准确率。(2)研究更好的规则获取方法，提高规则的精度和覆盖率。(3)根据抽取的正确结果，统计该结果前后词对该结果影响，提高抽取的准确率。

随着人物属性抽取这一技术的成熟，我们将可以在人物搜索这一领域提供智能的搜索服务。

参 考 文 献

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

[2] 王全剑, 李芳. 基于 Wikipedia 的人名简历信息抽取 [J]. 计算机应用与软件, 2011, 28(7): 170-174.

[3] 丁君军, 郑彦宁, 化柏林. 基于规则的学术概念属性抽取 [J]. 情报理论与实践, 2011, 34(12): 10-15.

[4] 车万翔, 刘挺, 李生. 实体关系自动抽取 [J]. 中文信息学报, 2005, 19(2): 1-6.

[5] 陆科进, 李新颖. 基于 Ontology 的文本信息抽取 [J]. 计算机应用研究, 2003: 46-49.

[6] 化柏林, 郭江. 基于规则的高校实验室 Web 信息抽取的系统设计与实现 [J]. 情报分析与研究, 2009(10): 62-66.

[7] 于满泉. 面向人物追踪的知识挖掘研究 [D]. 北京: 中国科学院计算技术研究所, 2006.

[8] Chen J X, Ji D H, Tan C L, et al. Unsupervised feature selectionfor relation extraction [J]. In Proceedings of IJCNLP, 2005: 262-267.