基于规则的人物信息抽取算法的研究

乔 磊[1,2],李存华[2],仲兆满[2],王 俊[2],刘冬冬[2]

(1. 中国矿业大学计算机科学与技术学院,江苏 徐州 221116)
(2. 淮海工学院计算机工程学院,江苏 连云港 222005)

[摘要] 随着互联网的快速发展,信息也呈爆炸式增长,如何从海量的文本信息中获取所需的信息成为当今一门重要的课题. 检索、分类、抽取等文本信息处理技术取得了长足发展,但面向人物属性的自动信息提取却没有引起人们的重视. 基于规则的人物信息抽取算法,首先对需要抽取的信息进行规则描述,重点是时间、地点、籍贯等信息. 在规则的基础上,研究开发人物信息抽取系统,最终实现了半结构化人物属性信息的自动提取.

[关键词] 文本信息抽取,人物信息抽取,人物属性规则,抽取算法
[中图分类号] TP391.1 [文献标志码] A [文章编号] 1001-4616(2012)04-0134-06

Research on People's Information Extraction Based on Rules

Qiao Lei[1,2],Li Cunhua[2],Zhong Zhaoman[2],Wang Jun[2],Liu Dongdong[2]

(1. School of Computer Science and Technology,China University of Mining and Technology,Xuzhou 221116,China)
(2. School of Computer Engineering,Huaihai Institute of Technology,Lianyungang 222005,China)

Abstract: With the rapid development of internet information with the explosive growth,how to obtain the required information from the vast amounts of text information is becoming an important issue today. Text retrieval,classification,extraction and other information processing technology has made considerable progress,but the automatic information extraction for character attributes did not cause people's attention. Rule-based character information extraction algorithms need to first extract the information on the rules described,with emphasis on time,place,event and other information. In the rule,based on research and development of character information extraction system,ultimately the character of semi-structured information automatically extracted.

Key words: text information extraction,people's information extraction,rules of People's attributes,extraction algorithm

在海量互联网信息中,人物信息也呈几何式增长,但总是数据丰富而信息贫乏. 当今人们获取信息的主要来源仍然是文本类型数据,如何对海量的人物文本信息进行有效的提取逐渐成为了人们关心的热点问题[1]. 传统的人工统计方法对人物信息的提取虽然有着准确度高的特点,但是低效、费时的缺点已不能满足当今信息时代的发展要求. 由于人物属性在文本中表现的复杂性、多样性,基于机器学习的方法也很难获得理想的结果.

在总结归纳人物属性在文本中的表现规律、制定详细规则的基础上,可以准确地从自由文本中提取出人物的一些基本属性,如姓名、性别、出生日期、政治面貌等信息. 本文以科学家领域的人物属性抽取为研究对象,包括出生日期、民族、性别、担任职务等属性的抽取.

人物信息抽取是人物传记和人物搜索引擎的基础工作,直接影响了人物传记和人物搜索引擎往"专、精、深"的方向发展.

1 研究现状

已有的研究工作与人物属性抽取相关的主要是人物传记和人物搜索引擎方面的研究. 人物传记不同于摘要,摘要侧重于从文本中摘录重要的信息,覆盖面全,而人物传记对摘录的信息针对性较强,只对人物

收稿日期:2012-07-10.
通讯联系人:乔磊,硕士研究生,研究方向:人工智能. E-mail:qiaony@163.com

相关的信息进行摘录,覆盖面较低. 从 1958 年,美国 IBM 公司的 H. P. Luhn [2] 开创了自动摘要至今,自动摘要已有 50 余年的历史,而最早是由 Schiffman [3] 在 2001 年提出的人物传记,人物传记可以从文献中提取人物性别、出生日期、学历等基本信息,最终形成传记性文字. 由于人物信息模式的相对固定,一些研究者构建了人物本体,用于抽取人物信息. 2007 年 Han 和 Park 等 [4] 利用 OWL 本体描述语言对人物信息建立事件本体,建立本体时把人物信息分为固定的和可变的,对人物的事件描述要素主要有: 人物、时间、地点、内容等,然后对人物基本信息和主要事件进行抽取.

面对互联网庞杂的数据,人们对于信息的准确性要求越来越高,人物搜索引擎作为新生事物,正处于兴起阶段,直到 2008 年才有成型的人物搜索引擎面世 [5]. 英文人物搜索引擎比较成熟的有雅虎人物搜索和微软的人立方等. 中文人物搜索引擎主要有优库、中文雅虎搜索引擎,优库主要是从网络资源中收集人物信息,包括姓名、性别、生日、身高、电话等个人基本信息; 中文雅虎人物搜索和微软的人立方主要是对人与人之间的关系的抽取,侧重于知名人士 [6],其中人立方抽取网页中人物的准确率是 97%.

另外,Zhong 等 [7] 研究了利用人物事件图从单文档中识别重要人物,在分析基于事件的文本摘要和人物传记对人物的研究存在不足的基础上,提出一种从单文档中识别重要人物的方法.

2　基于规则的人物信息抽取

2.1　基本框架

主要体系结构如图 1 所示.

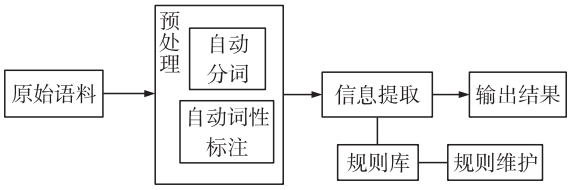

图 1　主要体系结构

Fig. 1　The main architecture

图 1 所示的体系结构包括几个部分: 1. 人物语料收集: 在 WEB 网上大量搜集一些特定人物的语料,如化学家、数学家、物理学家等,其中这些语料中包括人物的姓名、出生、职位、党派、职务等基本属性. 先对这些语料进行手工整理分类,一方面掌握了解一般文档对于人物属性信息描述的一般特征; 另一方面为以后对系统准确性测试作准备. 2. 构建人物属性信息规则: 归纳整理这些语料中属性一般的特征,如出生日期多出现在姓名之后、性别附近一定有姓名等特征,结合分词工具对文档进行自动分词,针对这些特征分别制定出相应的语法规则,在对人物属性提取的过程中分别运用这些规则,从而更好地把握所需要属性的准确性. 3. 提取人物信息: 根据制订的提取规则对文本所提取出来的人物属性进行有规律的组合.

2.2　人物属性信息的规则制订

规则对本文提出的人物信息抽取非常重要,它质量的好坏直接影响到信息抽取的效果. 在搜集语料的过程中,我们主要以科学家为例,搜集了 200 篇关于 20 位科学家的语料,在研究人物属性中我们发现,在人物属性的表述中有很多的相似性,根据这些相似性并结合分词工具,对文本的词性标注构建正则表达式.

2.2.1　文本词性标注

系统中采用的分词工具为中科院的 ICTCLAS,此工具能对汉语文本进行切分并标注词性. ICTCLAS 1.0(Institute of Computing Technology,Chinese Lexical Analysis System) 是中科院计算所开发的开源的汉语词法分析系统,该系统的功能有: 中文分词、词性标注和未登录词识别. 分词正确率高达 97.58%,基于角色标注的未登录词识别能取得高于 90% 召回率,其中,中国人名的识别召回率接近 98% [8].

表 1　常用词性

Table 1　Common part of speech

词性标注	n	nr	ns	t	f	v	m	q	p	c	w	nt	e
中文含义	名词	人名	地名	时间词	方位词	动词	数词	量词	介词	连词	标点	机构团体	叹词

2.2.2　构建正则表达式

正则表达式[9]是一种可以用于模式匹配和替换的规范,一个正则表达式就是由普通的字符(例如字符a到z)以及特殊字符(元字符)组成的文字模式,它用以描述在查找文字主体时待匹配的一个或多个字符串.正则表达式作为一个模板,将某个字符模式与所搜索的字符串进行匹配.

表2　正则表达式语法

Table 2　Regular expressions grammar

正则表达式语法	解释
\s	空白符号[\t\n\x0B\f\r]
\S	非空白符号[^\t\n\x0B\f\r]
.*?	0个以上的字符.

以下就以出生年月、籍贯、政治面貌为例:

1)通过对最常用的一些人物出生时间的表述研究,经过实验总结出了出生年月的正则表达式,如表3.

表3　出生年月正则表达式

Table 3　Date of birth regular expressions

正则表达式	语料举例
名字前缀类	
\S+/nr\s出生+.*?+/t+/nr+*?	李明1923年3月4日出生于(在)北京的一个普通家庭
\S+/nr\s+/w+/m+.*?/m+/w+/ns\s+人|	李明(1923--1979),北京人
\S+/nr\s+/w+/m+.*?+/w+/ns\s+人	李明(1934--),北京人
\S+/nr\s生于/t\s+/ns\s+人	李明生于1934年3月4日,北京人
\S+/nr\s于+/t\s+/ns\s+出生	李明于1934年出生于北京
\S+/nr\s生于/t+\s+籍贯+/ns+.*?+人	李明生于1934年3月4日,籍贯北京人士
\S+/nr\s性别+/n+出生于/t\s+/ns+人	李明,男,出生于1934年3月4日,北京人
\S+/nr+/w+/t\s+.6?+生+/w+/ns\s+人|	李明,1934年生,北京人
\S+/nr+/w+/t\s+.6?+出生+/w+/ns\s+人	李明,1934年出生,北京人
\S+/nr\s诞生.*?+/t\s+/ns\s+人|	李明诞生(于)1934年
\S+/nr+.*?+生辰+.*?+/t\s	李明,生辰,1934年3月4日
\S+/nr+.*?+出生日期+/t\s	李明,出生日期:1934年3月4日
后缀类	
\S+出生于+/t+的+/nr|	出生于1934年的李明
\S+出生于+/t+/ns+的+/nr	出生于1934年北平的李明
\S+生在+/ns+的+/nr	生在1934年的李明

以上是出生年月的正则表达式,可以总结出,名字前缀类可分为3类:

第一类:姓名+出生年月,

第二类:姓名+性别\|民族+出生年月,

第三类:姓名(出生年月-出生年月).

名字后缀类有2类:

第一类:出生年月+姓名,

第二类:出生地+姓名.

2)籍贯正则表达式

通过对大量文本的分析,可以对于关键字如"籍贯""祖籍""祖上""**人氏""***人"等对文本进行抽取.如李明,北京人,则提取籍贯为"北京"[10],见表4.

表4　籍贯正则表达式

Table 4　Birthplace regular expressions

正则表达式	语料举例
\S+籍贯.*?+/ns	籍贯北京
\S+原籍.*?+/ns	原籍北京
\S+祖籍+*?+/ns|\S+祖上.*?+/ns	祖籍北京|祖上北京
\S+/nr+.*?+/ns+人	李明北京人
\S+/nr+.*?+ns人氏|\S+/nr+.*?+ns人氏	李明北京人氏|北京人氏
\S+/nr+祖居+/nr	李明祖居北京

3)通过对人物属性信息的分析,在政治面貌方面总结出如下正则表达式(表5).

表5　政治面貌正则表达式

Table 5　Political status regular expressions

正则表达式	语料举例
\S + /nr\s + 中共党员	\S + /nr\s + 党员	李明,中共党员;李明,党员
\S + /nr + /t + .* ? 加入.* ? + 中国共产党	李明1988年光荣的加入了中国共产党
\S + /nr + /t + 加入 + .* ? + 中国民主党派名称	李明1988年3月4日加入九三学社
\S + /nr + .* ? + 共产主义战士	李明是一名坚定的共产主义战士
\S + /nr + 中国民主党派名称 + 成员	李明 九三学社成员
\S + /nr + .* ? + 人大代表	李明在2000年当选人大代表
\S + /nr + .* ? + 入党	李明于2000年入党

对于人物的其他属性,主要运用的还是归纳总结的方法,根据不同的语言表述习惯,得到相应的表达范式.

其中对中国民主党派名称,本文构建了中国民主党派库,代替了ICTCLAS中词性为"/e"的叹词.

2.3　基于规则人物属性信息的提取算法

本文中的抽取规则主要是根据正则表达式进行快速定位所要抽取的信息块并快速抽取信息.本文对收集的200篇关于科学家简介的文本进行了统计分析,设待识别词与正则表达式关键词距离为k,从200篇文本中查找到的词在待识别词的后面、前面以及k距离.

我们以出生年月的识别为例,其关键词是正则表达式中的如"出生"、"生于"等,其中在200篇语料中,出生时间出现了165次,见表6所示.

表6　出生年月与关键词之间的距离统计

Table 6　Distance statistics between date of birth and keywords

距离	出生时间出现在关键词后面的个数	出生时间出现在关键词前面的个数
K=1	29(17.58%)	3(1.82%)
K=2	47(28.48%)	2(1.21%)
K=3	34(20.61%)	2(1.21%)
K=4	16(9.70%)	1(0.61%)
K=5,6,7	25(15.15%)	0
K>8	3(3.64%)	0
个数汇总	157	8

从表6可见,对于出生日期,其出现在关键词之前的比例为:157/165 = 95.15%,其出现在关键词之后的比例为:8/165 = 4.85%.可以看出,出生日期出现在关键词后面占大多数,这说明在查找名词时应该先向后查找,如果找不到,再向前查找.对名词出现在关键词后面的情况,取$k=8$时,包含了95.15%的覆盖率,所以在本文中,从关键词的位置向后查找时的距离定为8($k=8$).对于出生日期出现在关键词的前面的情况,取$k=4$时,包含了100%的覆盖率,所以在本文中,从关键词的位置向前查找时的距离定为4($k=4$).

(1)算法一:出生年月

根据表3~4可知,出生日期出现在正则表达式关键词后面的概率远大于出现在后面的概率,故若发现触发词,先往后找词性为"/t"且中心为"月"、"日"的长度为词.若没有发现词性为"/t"的词,就向前查找4个字符内的词,进行分析后提取.

(2)算法二:籍贯

根据正则表达式中的关键词往后顺序6个字符,主要查找词性为"/ns"、"/f"的词.另外如果是姓名之后8个字符出现词性为"/ns"的词,也可判断这个词就是目标词.

(3)算法三:政治面貌

由于本文构建了中国党派库,代替了ICTCLAS中词性为"/e"的叹词,根据正则表达式可知,当发现正则表达式中的关键词时,可以往后顺序查找6个字符,查找词性为"/e"词.如果没发现词性为"/e"的词,再查找正则表达式中关键词如"人大代表"、"入党"等词.

3　实验结果与分析

在互联网上搜索了150篇介绍科学家的文章,并对搜集的语料中的人物属性信息进行手工整理、分析,对人物属性进行抽取,制定标准答案.现以科学家王选的简介为历,并使用XML语言描述答案.

例如:王选(1937年2月5日~2006年2月13日),男,汉族,江苏无锡人,生长于上海,九三学社成员,九三学社副主席.1958年9月参加工作,北京大学数学力学系计算数学专业毕业,大学学历,教授,中

国科学院院士,中国工程院院士,曾任北京大学计算机研究所所长,1996年~1998年当选九三学社中央副主席,任方正控股有限公司董事局主席,2003年3月在全国政协十届一次会议上当选为第十届全国政协副主席.

<attribute>
<brith>1937年2月5日</brith>
<name>王选</name>
<sex>男</sex>
<nation>汉族</nation>
<record>大学</record>
<politics>九三学社</politics>
<position>
北京大学计算机研究所所长
九三学社中央副主席
方正控股有限公司董事局主席
第十届全国政协副主席
</position>
</attribute>

对于属性识别的性能评价,采用

$$F-Score=\frac{2PR}{P+R}.$$

其中P为准确率,R为召回率,其定义分别为

$$P=\frac{M}{N},$$

$$R=\frac{M}{C}.$$

其中:M为找到时的正确的属性个数;N为找到属性的总数;C为正确的属性总数.

表7 人物属性抽取实验结果

Table 7 People's information extraction experiment result

人物属性	答案总数	结果总数数	正确数	准确率(P)	召回率(R)	F值
出生年月	132	136	112	0.82	0.85	0.84
性别	136	127	113	0.89	0.83	0.86
学历	126	168	115	0.68	0.91	0.78
民族	134	126	122	0.97	0.91	0.94
籍贯	129	136	114	0.84	0.88	0.86
职务	562	423	410	0.97	0.73	0.83
政治面貌	136	125	118	0.94	0.87	0.90

通过实验结果可知,基于规则的人物属性信息抽取对人物的性别、政治面貌及民族识别率还是比较高的,这也是因为人们对于性别、民族的表述结构比较单一,所以实验结果还是比较令人满意的.

对于人物职务、学历的识别相对不高的原因是:人物可能会担任多个职务,人物也可能会经过几个学习阶段,会出现多个关键词对最高学历判断的一定影响,加上人们对于职务、学历表述的多样性,所以还不能完全对职务及学历抽取,实验结果还有待提高.

4 结束语

在面对信息时代的海量信息时,信息抽取已变得非常重要.本文针对人物文本语料的基本属性提出了基于规则的人物属性信息的提取方法,在理论和实现方法上进行了深入的分析和探讨,分析了在提取人物属性信息中的一些触发词、特征词.制订了相应的规则,明确了人物信息的抽取内容,是人物属性和人物事件抽取的基础.对人物的基本属性信息,如:姓名、出生年月、学历、民族、籍贯、政治面貌、职务等做出了比

较深入的分析.本文只是对人物传记摘要研究的初步探索,还只是比较简单地对人物基本属性信息的抽取,系统还不完善.相信随着科技的进步,人物传记及人物搜索引擎这些领域一定会有更广阔的未来.

参考文献

[1] 易平,刘宗田,周文.人物传记研究综述[J].计算机工程与设计,2009,30(14):3426-3428.

[2] Luhn H P. The automatic creation of literature abstracts [J]. IBM Journal of Research Development,1958,2(2):159.

[3] Schiffman B,Mani I,Concepcion K. Producing biographical summaries: combining linguistic knowledge with corpus statistics [C]//Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics(ACL'2001). New Brunswick, New Jersey: Association for Computational Linguistics,2001:450-457.

[4] Han Y J,Park S Y,Park S B,et al. Reconstruction of people information based on an event ontology [C]//Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. Beijing,2007:446-451.

[5] 任宁.大规模真实文本中的人物职衔信息抽取研究[D].北京:北京语言大学信息科学学院,2008:4-7.

[6] 周婷.异构信息源的领域人物信息抽取研究[D].北京:哈尔滨工业大学计算机科学与技术学院,2010:6.

[7] Zhong Z M,Liu Z T,Li C H,et al. Identifying key people from a single document using people event map [J]. Journal of Computational Information Systems,2010,6(1):17-23.

[8] Hayneschan,W-OU,Anders,et al. ICTCLAS [EB/OL]. [2012-08-29]. http://baike.baidu.com/view/1215398.htm.

[9] 邓凯元,姜磊.正则表达式匹配引擎性能分析[J].计算机与现代化,2011(7):105-110.

[10] 颜伟王,洁尚英,宋柔.《中国大百科全书》人物传记知识提取加工规范语言[C]//全国第七届计算语言学联合学术会议论文集.哈尔滨,2003.

责任编辑:黄　敏