基于序列模式挖掘的人物关系识别*

李丹　罗智勇

北京语言大学语言信息处理研究所 北京 100083

Email: {lidan, luo_zy}@blcu.edu.cn

摘要： 命名实体关系抽取是信息抽取领域中的重要研究课题。本文利用序列模式挖掘方法，从大规模生语料中自动提取表达人物关系的序列模式，用于人物实例关系抽取；为了避免数据稀疏问题而导致模式遗漏，我们将具有相同文本表达模式的特征词语进行聚类，以提高关系模式的覆盖率；同时给出了一种序列模式评估方法，将准确率高的模式进行优先匹配，以提高人物关系识别的正确率。实验结果表明，人物关系自动识别的正确率和召回率分别达到了 84%和 75.2%。

关键字： 信息抽取，人物信息检索，序列模式，模式挖掘控制

Persons' Relationship Recognition Based on Sequential Pattern Mining

LI Dan　LUO Zhiyong

Center for Language Information Processing, Beijing Language and Culture University, Beijing 100083

Email: {lidan, luo_zy}@blcu.edu.cn

Abstract: Discovering relations between named entities in the field of information extraction is an important research topic. In this paper, the patterns describing the persons' relationship are extracted automatically by the mean of sequential pattern mining. In order to avoid sparse data problem, the similar words which have the same expression pattern characteristic in the text will be clustered and that will improve the coverage of patterns. The method of assessing the patterns is given. The high accurate pattern will be matched with priority to improve the accurate of persons' relationship recognition. The experimental results show that our approach is suitable for the information extraction on people relationship. The accuracy reaches 84% and the recall rate reaches 75.2%.

Keywords: information extraction, people relationship research, sequential pattern data mining, pattern refining

1　引言

近二十年来，随着计算机技术和互联网技术（Internet）的发展，海量文本的储存、传送已经变得十分便捷。同时用户不再满足于文档相关性检索，而转向结构化、细颗粒度、精准化的信息检索需求。这一变化进一步促进了信息抽取（Information Extraction）技术的发展。本文的工作关注于人物实例关系抽取研究，研究成果可以服务于人物信息检索系统和自动问答系统等等。

信息抽取系统任务是将无结构的文本转化为半结构化或者结构化的信息，并以数据库方式保存，供用户查询处理。到目前为止，信息抽取领域已有许多关系抽取方法被应用在各种实验系统当中。这些方法技术基本可以归纳为：

1)基于模式匹配的关系抽取。最初关系模式的建立需要依靠语言学家对抽取任务涉及的领域语料进行深入分析，手工编织关系模式，应用成本很高；另一方面，该方法可移植性差，无法应

*基金项目：教育部科学技术研究重点项目（107017）的资助；

作者简介：李丹，女，1985 年生，硕士研究生，主要研究方向自然语言处理；罗智勇，男，1975 年生，博士，讲师，主要研究方向为计算语言学。

用在新领域的关系抽取。针对这一问题 Douglas E. Appelt 等人[3]在 MUC - 6 上提出的 FASTUS 抽取系统中,通过引入“宏” 的概念将各种领域依赖规则以一种具有扩展性的、 通用方式表达。用户只需要修改相应“宏” 中的参数设置,就可以快速配置好特定领域任务的关系模式规则。Roman Yangarber 等人[4]在 MUC - 7 上提出的 Proteus 抽取系统采用了基于样本泛化的关系抽取模式构建方法。这些改进方法为基于总结规则节约了大量的时间。

2)基于词典驱动的关系抽取。Chinatsu Aone 等人[5]在 MUC - 7 上提出了一个快速、 灵巧的大规模事件和关系抽取系统 REES。此方法的缺点是，它只能识别以动词为中心词的关系，而对名词同位语之类的关系抽取就很难实现。

3)基于机器学习的关系抽取。车万翔等人[6]使用两种基于特征向量的机器学习算法 Winnow 和 SVM 进行实体关系抽取。除此之外刘克彬等人[7]实现了基于核函数的中文实体关系自动抽取系统，应用改进的语义序列核函数，结合 KNN 机器学习算法构造分类器来分类并标注关系的类型。而这些算法的先决条件是需要预先通过对一个特定领域手工标引的数据集进行学习以获取领域内关系类型的各项特征。针对这一问题，近年来出现很多弱监督学习的抽取方法。这些方法都是根据 Zhu Zhang[8]提出的基于 SVM 的弱监督关系分类系统应用 SVM 算法进行关系抽取。其代表性工作有：张素香[9]和何婷婷[10]等基于种子扩展的方法来得到实体关系。这种算法有一定风险，在种子扩展或者模板生成的过程中没有任何人工干预，如果扩展的种子词或者模板不准确，则会导致最后结果准确率很低。

4)基于语料库的统计方法。该类方法主要通过串频统计策略和基于 n 元文法的统计语言模型方法，获取词语线性邻接特征和信息表达模式，为特定领域的信息抽取系统服务。代表性的研究工作有：Deepak[11]首先通过搜索引擎收集含有特定信息描述（出生日期[Birthday]、发明者[Inventor]等））的文本，然后采用串频统计方法，获取该类信息的描述模式用于自动问答（QA）系统的实现，并在 TREC 评测中取得优异成绩。但该方法考察的仅是信息表达的线性邻接模式，并没有考虑到隔距模式。

通过以上对目前技术方法的对比分析，本文利用序列模式挖掘[14]方法，该方法既不需要标注语料，又考虑了隔距模式，同时在生成的模式中，进行对模式进行评估，最终产生出准确率较高的模式。实验结果表明该方法效果较好。

2 人物关系逻辑分类与特征分析

人物关系是指人与人之间因为所处的社会位置和所承担的社会角色所带来的社会关系，为了便于理解，我们将人物实例之间的关系大致分为四个大类，每个大类细分为若干子类，如下表 1：

表 1 人物关系逻辑分类表

关系类别	关系子类
家庭关系	长辈关系、晚辈关系、兄弟姐妹关系、亲家关系、夫妻关系
工作关系	上司关系、下属关系、同事关系
朋友关系	好友关系、情侣关系
其他关系	师徒关系、邻居关系、同伙关系、敌对关系、师徒关系

在真实文本中，人物关系的描述通常具有一定的表达模式，根据对大量语料的统计分析，

大概有以下三个特征：

1) 人名与关系特征词语之间具有顺序性。如“李鹏夫人朱琳”、“邓小平与撒切尔夫人”。第一句中的人物关系是正确的，而第二句中的“夫人”并不是表示妻子这样的关系，而是一个后身份词。对于像“夫人”这样的可以表示后身份词的关系词出现在两个人名之后表示关系的准确率较低，也就是说序列“<人名><人名><夫人>”的准确率比序列“<人名>夫人<人名>”的准确率低。因此这种关系特征词语出现在人名之间的位置关系是需要考察的。

2) 人名、关系词语之间不但有次序关系，其间还可能有多个插入成分，呈多元隔距搭配模式。如：“国务院总理李鹏今天上午在中南海紫光阁与夫人朱琳会见.....”。在人名与人名之间，或者人名与关系词之间是否相邻或者相隔若干词语，是否有一些词语一定会出现在它们之间，这些问题都将体现在人物关系文本的表达模式上。

3) 表达同类人物关系的特征词语具有多样性。例如“妻子”关系的表述，可以使用“夫人”、“妻子”、“爱妻”、“贤内助”等等，其中有些特征词语使用非常普遍，而有些却十分罕见。

特征 1、2 表明要准确提取人物关系，需要鉴别人名和特征词语的顺序关系和远距离搭配关系，因此我们采用基于序列模式挖掘[14]的方法来生成人物关系的文本表达模式。特征 3 表明需要在生成模式之前，将表达同类人物关系的特征词语进行聚类，这样既可以解决数据稀疏的问题，同时也可以将人物关系的处理变得更为灵活，同样一种关系，可以有多种特征词语的表达方式，以提高关系模式的覆盖率。

3 基于序列模式挖掘的人物关系识别

生成人物关系模式的工作主要有四个部分：语料库预处理、人物关系的序列模式挖掘、模式评估以及人机交互对模式的筛选。

3.1 语料库预处理

建立词表 V={w1, w2,……,wn}，其中多数是普通词语，另外还可能是类别名，如人名、地名、机构名、数字和时间短语等等。同时通过对特征语料库的筛选与分析，得到关系特征词语库，其中包含 636 个关系特征词。如前所述，由于特征词语的多样性和数据稀疏问题，我们通过人机交互的方式，根据词语在大规模语料库中的上下文分布相似性，对这些特征词语进行聚类处理，减少低频特征词语相关表达模式的遗漏。但这些类别只是表征词语在文本表达方式的上的相似性，即：同一个特征词语聚类内的两个不同词语可能表达同种人物关系（如“夫妇”和“夫妻”都表达“夫妻”关系），也可能表达不同人物关系（如：“夫妇”和“师徒”），但“夫妇、夫妻、师徒”等属于复合关系词语，具有相似的文本表达方式。由于篇幅的原因，词语上下文分布相似性计算方法另文论述[13]。

3.2 前缀投影序列模式增长算法

1) 算法

通过前缀投影序列模式增长算法（PrefixSpan）[15]，序列模式的生成是一个前缀不断增长的过程，当前缀长度达到设定的序列模式长度范围，同时包含两个人名和一个关系词时，则当前前缀即为一个序列模式。此时可将当前前缀输出，则生成一条序列模式。

2) 模式限制条件

根据生成模式的初步实验，发现一些词符合最小支持度和最小转移概率的条件，但对于人物关系信息表达无关，如：日期时间，空格，标点符号，还有一些单字词，如“得”，“了”等。在生成模式时，忽略对这些词的计算，很大程度地降低时间复杂度。

3.3 模式评估

实际上，利用上述算法生成的模式集中，不同模式的使用频度和抽取的人物关系实例的正确率是不一致的。下面我们给出一个利用人物实例关系可信度来估计模式正确率的方法。

1) 人物实例关系可信度

为了准确计算关系实例的出现概率，我们将关系分为唯一关系和非唯一关系。

唯一关系的定义如下：对于人物库S，人物$a \in S$，人物$b \in S$，唯一关系R_u，如果存在aR_ub，则必不存在$\bar{a}R_ub$或者$aR_u\bar{b}$，而且必不存在$aR_u'b$（R_u'是R_u以外的其他唯一关系）。关系词库中，如“妻子”、“父亲”就属于唯一关系的关系词。

非唯一关系是除唯一关系以外的其他关系。

a. 唯一关系R_u实例的可信度计算方法

$$P(b \mid aR_u) = Max\{P(x \mid aR_u)\} = Max\{\frac{P(aR_ux)}{P(aR_u)}\}, \forall x \in S \text{。}$$

b. 非唯一关系R实例的可信度计算方法

非唯一关系不具有排他性，假设$\forall x \in S, aRx$从逻辑上均可以成立，而要判断其是否准确，需要得到关系R的可信度，即$P(R \mid ax) = \frac{P(aRx)}{P(ax)}$。

2）模式评估

一个模式对应了多条由它匹配出来的实例，模式的可信度取决于它匹配的实例的可信度值。假设，模式k匹配实例N个，每个实例i可信度值为S_i，则模式k可信度$P_k = \frac{\sum_{i=1}^{N} S_i}{N}$。

3.4 模式筛选

在生成的模式中，去掉可信度小于 0.5 的模式。另外再加入一些人工知识对模式进行进一步加工。说明如下：

1) 对模式中的通配符“.”做限定。当句子出现两个以上人名或多个关系词时，模式中的通配符“.”会匹配任意字符，包括人名和关系词。例如：“记者周效政报道，中国棋手秦侃滢战平

对手马里奇。”当规则“<人名>.*对手.*<人名>”匹配句子时，会提取“周效政——对手——马里奇”，把中间的“秦侃滢”当做“.*”给匹配了。所以限定通配符“.”不能为人名或者关系词。

2) 对模式中人名与关系词的距离做限定。在实验过程中，不断找出一些匹配错误的句子，发现人名与关系词之间的距离非常重要，一般关系词后面如果有人名即，顺序为“<人名>…<关系词>…<人名>”或“<关系词>…<人名>…<人名>”的句子，关系词后面的人名会紧跟在关系词后，中间不会有其他词出现。

4 实验结果及评测

4.1 准确率的评测

本实验准确率的评估为：正确率=正确实例个数/随机抽取人名关系个数。

根据第2节给出的算法，用一个近20万词的词表对8年《人民日报》语料进行训练生成模式，对模式进行封闭式测试，一共提取人物关系4717个，随机抽取100个人名关系查看其结果；同时实际抽取了另外3年《人民日报》语料作为评测语料，一共提取人物关系1414个，随机抽取50个人名关系查看其结果。为了对评测结果有直观的认识，直接从训练语料和评测语料中提取包含两个人名与一个关系词的句子，然后分别随机抽取100个、50个人物关系，将其正确率作为Baseline，与本文的实验结果进行对比。

评测结果如下表2：

表2 准确率评测结果

	序列挖掘	直接提取（Baseline）
8年语料封闭测试	82%	38%
3年语料开放测试	84%	46%

4.2 召回率的评测

本实验召回率的评估为：召回率=正确实例个数/文本正确人名关系个数。

召回率的测试语料是人民日报2000年12月的文章，挑出其中包含两个或两个以上人名标注，和关系词的句子一共224条，人工筛选包含正确人名关系的句子101条。

同时用本文的方法提取测试语料中人物关系，一共匹配出86条，人工筛选正确的人物关系一共76条，召回率=76/101=75.2%.

4.3 实验结果对比

下面将其他方法与本文方法实验结果进行对比，其他方法均取各自F值最高时的结果。

表3 实验结果对比

抽取方法	序列挖掘	Winnow	SVM	Bootstrapping
准确率	83%	74.75%	76.13%	82.92%
召回率	75.2%	71.69%	70.18%	73.8%

5 结语

本文针对于人物关系识别，采用了序列模式挖掘的方法来自动生成出描述人物关系的模式，对人物关系识别做了基础的探索性实验，在大规模语料中运用这些模式准确率较高，取得较好的效果。但仍然还有一些问题有待进一步解决，例如：一人多名“江泽民”与“江主席”，如何将其识别为同一个人。以及像“翻译”，这样既可以表示关系，也可以表示动作的多义词，如何进行歧义的消解，并且能够在句法、语义分析等深层次处理技术上获得更高的准确性。这些问题都是需要进一步研究探讨的问题。另外在这基础上应该有所扩展，可以将人物之间的关系网络化，这要求建立更完善的人物关系库。所以，在今后的工作中，我们将进一步细化人物关系识别的颗粒度，以达到更好的效果。

参 考 文 献

[1] M.E.Califf. Relational Learning Techniques for Natural Language Information Extraction [D]. Univ. of Texas, 1998.

[2] MUC.1987-1998.The nist MUC website: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[3] Appelt D E, Hobbs J R, Bear J, et al . SR I International FASTUS System: MUC - 6 Test Results and Analysis[C]. In: Proceedings of the 6th Message Under standing Conference (MUC - 6) , 1995: 237- 248.

[4] Roman Y, Grishman R. NYU: Description of the Proteus/PET System as Used for MUC - 7 ST[C]. In: Proceedings of the 6th Message Understanding Conference (MUC - 7), 1998..Aone C, Ramos SantacruzM. Rees: A large - scale relation and event extraction system[C]. In: Proc of the 6th Applied Natural Language Processing Conference, New York, 2000: 76 - 83.

[5] Aone C, Ramos SantacruzM. Rees: A large - scale relation and event extraction system[C]. In: Proc of the 6th Applied Natural Language Processing Conference, New York, 2000: 76 - 83.

[6] 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

[7] 刘克彬,李芳,刘磊,基于核函数中文关系自动抽取系统的实现[J],计算机研究与发展,2007,44(8):1406-1411.

[8] Zhu Z . Weakly - supervised Relation Classification for Information Extraction[C]. I n: Proceedings of the Thirteenth ACM conference on Information and Know ledge Management, Washington D. C. ,2004: 581 - 588.

[9] 张素香,李蕾等,基于 Boost Strapping 的中文实体关系自动生成[J],微电子学与计算机,2006,23(12):15-18.

[10] 何婷婷,徐超等,基于种子自扩展的命名实体关系抽取方法[J],计算机工程,2006,32(21):183-184.

[11] Deepak Ravichandran and Eduard Hovy. Learning Surface Text Patterns for a Question Answering System[C].ACL, Philadelphia, July 2002:41-47.

[12] 于琨,管刚等,基于双层级联文本分类的建立信息抽取[J].中文信息学报，第 20 卷 第 1 期.

[13] 罗智勇,宋柔,相似词及其在计算机辅助校对中的应用[C],《自然语言理解与大规模内容计算》，全国第八届计算语言学联合学术会议(JSCL-2005)论文集, 2005.8

[14] Jiawei Han, Micheline Kamber 著,范明,孟小峰译,数据挖掘概念与技术[M],机械工业出版社,2006:325-336.

[15] J.Pei, J.Han. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth[C]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(11), 1424- 1440.