基于 NNV 关联规则的非分类关系提取方法及其应用研究*

王 红[1,2], 高斯婷[1†], 潘振杰[1], 肖志伟[1]
(1. 中国民航大学 计算机科学与技术学院, 天津 300300; 2. 天津大学 管理学院, 天津 300072)

摘 要: 针对传统的非分类关系提取方法无法获得非分类关系的名称的不足,提出基于 NNV(noun-noun-verb,名词名词动词)关联规则的非分类关系提取方法。给出 NNV 关联规则的相关概念及方法的实现过程,提取了民航突发事件应急管理领域本体中的非分类关系,完善了民航突发事件应急管理领域本体。与传统的非分类关系提取方法相比,有效获取了非分类关系的名称,保证了结果的准确率和召回率。

关键词: 非分类关系; 关联规则; 民航突发事件; 领域本体
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)10-3665-04
doi:10.3969/j.issn.1001-3695.2012.10.016

Application and research of non-taxonomic relation extraction method based on NNV association rule

WANG Hong[1,2], GAO Si-ting[1†], PAN Zhen-jie[1], XIAO Zhi-wei[1]
(1. *School of Computer Science & Technology, Civil Aviation University of China, Tianjin* 300300, *China*; 2. *School of Management, Tianjin University, Tianjin* 300072, *China*)

Abstract: To solve the problem that traditional methods can't get the tag of non-taxonomic relation, this paper proposed a non-taxonomic relation extraction method, based on noun-noun-verb association rule in which the correlative concepts were defined by combining the nature language processing and association rule. The method was applied to civil aviation emergency domain. Compared with traditional methods, the method finds the name of non-taxonomic relation effectively and the result represents a better precision and recall.

Key words: non-taxonomic relation; association rule; civil aviation emergency; domain ontology

0 引言

自 20 世纪 90 年代以来,随着知识共享、信息集成、语义 Web 和 Web 服务等技术的快速发展,本体学习方法逐渐成为研究的热点。近年来,本体学习在概念提取和分类关系提取上的研究进展较多,而非分类关系的研究还处于探索阶段,针对实际应用的研究则更少。

目前,提取非分类关系主要有基于词典的方法、基于模式匹配的方法和基于关联规则的方法[1,2]。基于词典的方法抽取出的关系必须是 WordNet 中已经存在的关系,并仅能从 WordNet 中抽取同义、反义和部分/整体这几种关系。基于模式匹配的方法主要利用预定义的非分类关系模式提取非分类关系,提取结果通常都是概念间的"匿名"关系。基于关联规则的方法[2]是依据"如果概念 C_i 出现,概念 C_j 也会出现,则概念 C_i 和 C_j 存在某种关系"的原理,通过频繁挖掘的关联规则思想来判断概念 C_i 和 C_j 是否存在非分类关系,同样无法获得概念间的具体名称[3~5]。

本文在深入分析现有非分类关系提取方法的基础上,将关联规则与自然语言处理方法相结合,提出了一种 NNV 关联规则法,并将其应用在民航突发事件应急管理领域本体[6]的构建之中,为本体学习的深入研究提供了良好的支持。

1 NNV 关联规则

自然语言处理方法一般都将动词视做句子中最能表达关系的信息[7,8]。例如,语句"教师教授课程",存在"教师"和"课程"两个概念以及"教授"的关系,可表示为 RDF 三元组(教师,教授,课程),对应于本体中的非分类关系为"教授(教师,课程)"。由此发现概念和非分类关系的三元组主要是(主语,谓语,宾语)的形式,与句子中的"主谓宾"形式类似。依照本体构建标准"名词是类名的基础,动词或者动词短语是属性名的基础",本文将句子中名词(noun)或名词短语的主语和宾语作为两个概念,动词(verb)或动词短语的谓语作为概念间的非分类关系,改进关联规则法,提出 NNV 关联规则法,将概念和非分类关系表示成规则 $(C_i, C_j) \Rightarrow R_t$ 的形式进行非分类关系的提取。

若概念 C_i 和概念 C_j 具有关系 R_t,则可以通过设置支持度

收稿日期: 2012-03-16; **修回日期**: 2012-04-26 **基金项目**: 国家自然科学基金委员会与中国民用航空局联合资助项目(61079007)

作者简介: 王红(1963-),女,重庆人,教授,主要研究方向为本体技术、数据挖掘与智能信息处理;高斯婷(1989-),女(通信作者),硕士,主要研究方向为本体学习、知识工程与语义网技术(gaositing 890213@ yahoo.com.cn);潘振杰(1987-),男,硕士,主要研究方向为语义网框架、语义检索;肖志伟(1988-),男,硕士研究生,主要研究方向为语义检索与应用.

和置信度来挖掘出满足规则$(C_i, C_j) \Rightarrow R_t$的概念对和非分类关系。本文将此种方法称为 NNV 关联规则法

定义 1 NNV 关联规则。设 $I = \{N_1, N_2, V_1, \cdots, N_i, N_j, V_i, \cdots, N_r, V_z\}$是项的集合,其中 N_i 为名词或名词性短语,V_j 为动词或动词性短语,给定一个事务集 $T = \{t_i \mid i = 1, 2, \cdots, n\}$, $t_i = \{N_k, \cdots, N_t, V_t, \cdots, N_s, V_p \mid N_t \in I, V_t \in I\}$,将满足最小 NNV 支持度阈值 min_NNV_sup 和最小 NNV 置信度阈值 min_NNV_conf 的规则:

$$(N_i, N_j) \Rightarrow V_k (N_i \subset I, N_j \subset I, V_k \subset I, N_i \cap N_j \cap V_k = \varnothing)$$

称为 NNV 规则。

定义 2 NNV 支持度(NNV_support)。NNV 规则$(N_i, N_j) \Rightarrow V_k$ 的 NNV 支持度是指事务集中包含 $N_i \cup N_j \cup V_k$ 的百分比,即概率 $P(N_i \cup N_j \cup V_k)$,记为

$$\text{NNV\_support}((N_i, N_j) \Rightarrow V_k) = \text{NNV\_support}(N_i \cup N_j \cup V_k) = P(N_i \cup N_j \cup V_k)$$

定义 3 NNV 置信度(NNV_confidence)。NNV 规则$(N_i, N_j) \Rightarrow V_k$ 的 NNV 置信度是指事务集中包含 N_i 和 N_j 项并同时包含 V_k 项的事务的百分比,即条件概率 $P((N_i, N_j)/V_k)$,记为

$$\text{NNV\_confidence}((N_i, N_j) \Rightarrow V_k) = P((N_i, N_j)/V_k) = \frac{\text{NNV\_support}(N_i \cup N_j \cup V_k)}{\text{NNV\_support}(N_i \cup N_j)}$$

定义 4 频繁项集。NNV 规则$(N_i, N_j) \Rightarrow V_k$ 的频繁项集和传统关联规则的频繁项集具有类似的定义,即项集 I 的支持度满足定义的最小 NNV 支持度阈值 min_NNV_sup,则 I 为频繁项集。

2 基于 NNV 关联规则的非分类关系提取方法

为了获取概念间的非分类关系,可以将术语集(概念集和关系集)作为 NNV 关联规则项的集合,以领域文本集中的一个文档或一段文本或一句话作为一个事务,对事务中的每项进行词性标注,利用 NNV 关联规则法的思想,挖掘概念 C_i 和 C_j 间的非分类关系 R_t,即规则$(C_i, C_j) \Rightarrow R_t$。

非分类关系获取的具体流程包括构建领域词典、分词处理、过滤领域术语并提取概念和分类关系、获取事务集、发现概念对频繁二项集和提取非分类关系六大部分。

2.1 构建领域词典

由于文本中存在大量的领域词汇,分词工具无法识别,常常将领域词汇切分为散串,降低领域术语获取的准确性。因此首先需要构建领域词典,对领域中的专业术语进行归纳总结,为大量文本进行分词和提取领域术语做准备。领域词典包括领域词汇表和领域同义词关系表两个部分,构建流程如图 1 所示。

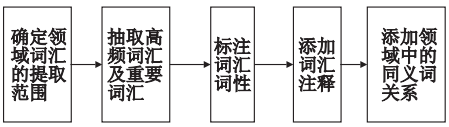

图1 领域词典的构建流程

2.2 分词处理

ICTCLAS 工具能够加载用户词典对领域文本进行分词,达到 98.45% 高精度。利用 ICTCLAS 对领域文本分词分"加载用户词典"和"分词"两个步骤进行。

1)加载用户词典

调用 ICTCLAS 工具的默认接口 ICTCLAS_ImportUserDict(byte[ ] sPath, eCodeType eCT),加载领域词典中的领域词汇表。

2)分词

调用 ICTCLAS 工具的默认文本分词接口 ICTCLAS_FileProcess(const char * sSrcFilename, eCodeType eCt, const char * sDsnFilename, int bPOStagged),完成对领域文本的分词处理。

2.3 过滤领域术语并提取概念和分类关系

对文本进行分词划分后,需要对文本进行处理,过滤得到领域术语。领域术语的获取包括如下四个步骤。

1)保留所需词性词汇

根据"名词是类名的基础,动词或者动词短语是属性名的基础"的原则,以及"一句话作为一个事务"的思想,对分词结果进行统计,保留文本中的名词或名词性短语、动词或动词性短语以及标点符号,初步过滤领域文本中的术语。

2)替代领域同义词

为了避免划分词汇后直接统计时出现数据稀疏的现象,需要将同义词词汇进行合并,利用公认的权威领域词汇来取代非权威词汇。利用 2.1 节中已构建的领域词典的同义词关系表,将同义词用统一的词汇表示。

3)GF/GL 权重过滤以及领域相关度和一致度过滤

利用 GF/GL 权值过滤[6]的方法以及领域相关度和一致度的过滤方法,提取领域术语集 S。

4)提取概念和分类关系

过滤的领域术语中,根据术语的词性选择名词或名词性短语作为领域概念,并采用层次聚类算法[9]来进行提取概念间的分类关系。

2.4 获取事务集

挖掘规则$(C_i, C_j) \Rightarrow R_t$ 首先需要获取项的集合以及事务的集合。项的集合即为所有领域术语。事务的集合获取的主要思想是:逐行扫描领域文本的句子,如果句子中存在领域术语(即概念和关系),则将此句作为一个事务,此句包含的所有领域术语作为此事务的项。

算法 1 事务集获取算法

输入:

$D = \{D_0, D_1, \cdots, D_i, D_j, \cdots, D_n\}$ 为领域文本;

n 为领域文本的总数;

D_{ij}为文本 D_i 的第 j 句话;

length(D_i)为领域文本 D_i 的句子总数;

$I = S = \{C_1, C_2, R_1, \cdots, C_i, C_j, R_i, \cdots, C_r, R_z\}$ 为项集;

S 为 2.3 节中的术语集。

输出:事务集 $T = \{t_k \mid k = 1, 2, \cdots, p\}$。

```
1.  k = 0, T = ∅
2.  for i = 0 to n do
3.    for j = 0 to length(D_i) do
4.    扫描 D_ij
5.    if D_ij 中包含元素 p_1 ∈ I, ···, p_t ∈ I then
6.    将集合 t_k = {p_1, ···, p_t} 放入 T 中并使 k 增 1
7.    endif
8.    endfor
```

9.　endfor

算法 1 详细地介绍了获取事务集的方法,为获取频繁项集做准备。

2.5　发现概念对频繁二项集

获取概念对频繁二项集也是为了挖掘两个概念是否有关系。利用 Aprior 性质“频繁项集的所有非空子集也必须是频繁的”,本文将已提取的概念作为频繁一项集,采用改进的频繁二项集连接步和频繁二项集剪枝步发现概念对频繁二项集。

1)频繁二项集连接步

将项集 $I=S=\{C_1,C_2,R_1,\cdots,C_i,C_j,R_i,\cdots,C_r,R_z\}$ 拆分为概念集合即名词性集合 $C=\{C_1,C_2,\cdots,C_i,\cdots,C_r\}$ 和关系集合即动词性集合 $R=\{R_1,R_2,\cdots,R_i,\cdots,R_z\}$。令频繁一项集 $L_1=C=\{C_1,C_2,\cdots,C_i,\cdots,C_r\}$。对于一项集 L_1 的元素,将 L_1 与自身连接产生候选二项集 Q_2。

2)频繁二项集剪枝步

扫描事务集 T,确定 Q_2 中所有候选的计数,计数值不小于最小 NNV 支持度阈值 min_NNV_sup,则保留候选项,从而得到频繁二项集 L_2。

频繁二项集 L_2 中的每项均为一对概念,即为有联系的概念对,还需要获得频繁二项集的概念对中的非分类关系名称,即发现概念对和非分类关系组成的频繁三项集。

2.6　提取非分类关系

发现概念对的频繁二项集后,存在关系的概念对被挖掘,还需发现概念对和关系连接组成的频繁三项集,并对其找出满足最小 NNV 置信度阈值 min_NNV_conf 的集合,即为最终的 NNV 关联规则。同 2.5 节,采用改进的频繁三项集连接步和 NNV 关联规则剪枝步进行。

1)频繁三项集连接步

将频繁二项集 L_2 中的每项和集合关系集合 $R=\{R_1,R_2\cdots,R_i,\cdots,R_z\}$ 进行连接,产生候选二项集 Q_3。

2)NNV 关联规则剪枝步

扫描事务集 T,确定 Q_3 中所有候选的计数,计数值不小于最小 NNV 支持度阈值 min_NNV_supp,则保留候选项,得到频繁三项集 L_3。根据定义 3 中的公式计算 NNV 置信度 NNV_conf,如果 NNV_conf 不小于最小 NNV 支持度阈值 min_NNV_conf,输出 NNV 规则。

最终得到的 NNV 关联规则 $(C_i,C_j)\Rightarrow R_t$ 代表概念 C_i 和概念 C_j 具有关系 R_t,即得到两个概念的非分类关系。

3　基于 NNV 方法的应用与实验结果分析

3.1　民航突发事件应急管理领域本体的非分类关系提取

以民航突发事件应急管理领域本体的构建为实验应用对象,获取《天津滨海国际机场应急救援计划手册 205》《首都机场应急救援手册》《中国民用航空应急管理规定》《民用运输机场应急救援规则》《世界民航事故数据库》等资源作为实验语料,处理得到 851 个文本文件。

采用关联规则提取概念间的关系时,尚无良好的计算方式设定置信度和支持度,因此 NNV 关联规则法采用逐步尝试的方法设定 min_NNV_sup 和 min_NNV_conf。当选取 min_NNV_sup = 0.5%,min_NNV_conf = 0.5% 时,实验结果最佳。

从以上文件中提取民航突发事件应急管理领域本体的非分类关系部分结果如表 1 所示。

表 1　民航突发事件应急管理领域本体的非分类关系表

概念 N	概念 N	关系 V	NNV_conf/%
民航突发事件	航空器与航空器相撞事故	包括	0.555 732
民航突发事件	应急预案	采用	1.655 422
爆炸品	民航恐怖事件	引发	1.346 210
航空器场内失事事故	机场	发生于	3.125 784
航空器跑道事故	跑道	发生于	3.205 881
机场应急救援指挥中心	应急预案	制定	2.468 575
机场应急救援指挥中心	机场应急救援部门	指挥	3.565 141
机场应急救援部门	消防部门	包括	2.578 114
机场应急救援部门	公安部门	包括	2.356 557
机场应急救援部门	医疗部门	包括	2.952 012
机场应急救援部门	驻场武警	包括	2.020 491
消防部门	机场应急救援指挥中心	通报	2.545 314
消防部门	机场消防线	聚集	1.965 026
公安部门	救援现场	封锁	1.364 447
公安部门	消防部门	协助	0.502 246
公安部门	警戒线	划定	1.532 316
公安部门	爆炸品	拆除	2.334 767
医疗部门	应急预案	启动	2.365 548
医疗部门	救护车	调度	2.544 325
医疗部门	第三中心医院	增援	0.654 527

将非分类关系加入原有民航突发事件应急管理领域本体中,形成完整的领域本体,本体图的部分结果如图 2 所示。其中,椭圆表示概念,实线代表概念间的分类关系,虚线代表概念间存在非分类关系;虚线上的字符代表实验中获得的非分类关系的名称。

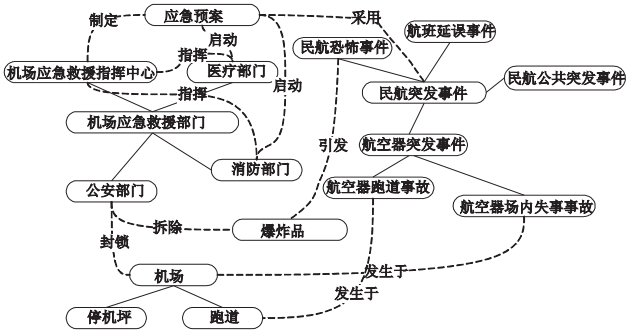

图2　民航突发事件应急管理领域本体图

由图 2 可以看出,概念间的关系不再是简单的分类关系,还包含非分类关系,从而构成了网状结构,完善了领域本体。例如,“航空器跑道事故”和“跑道”具有“发生于”的非分类关系,搜索“航空器跑道事故”,可以根据此完整的民航突发事件应急管理领域本体,利用语义检索技术得到事发地点为“跑道”;“公安部门”和“机场”具有“封锁”的非分类关系,搜索“公安部门”,可以检索出救援动作为“封锁”,动作对象为“机场”。通过非分类关系,可以更加全面地得到概念间的联系,使检索结果更加丰富,为民航突发事件应急管理提供了更完整的支持。

3.2　实验结果分析

为测验实验的稳定性,本文采用递增实验语料数量的方法进行分析,计算实验结果的准确率和召回率,如表 2 所示。

表 2 不同数量语料的非分类关系提取结果对比表

语料库中领域文本个数	正确非分类关系数	获取非分类关系数	所有非分类关系数	准确率 precision/%	召回率 recall/%
100	30	93	114	32.3	26.3
200	35	96	119	36.5	29.4
300	38	98	126	38.8	30.2
400	39	99	127	39.4	30.7
500	39	99	127	39.4	30.7
600	39	100	128	39.0	30.5
700	41	104	133	39.4	30.8
800	42	107	136	39.3	30.8

由表 2 可以看出,领域文本的数量从 100 增至 300 时,准确率和召回率逐步增加。随着文本数量增至 400 后,结果逐步趋于稳定,准确率和召回率分别趋近于 39.4% 和 30.7%。

传统的非分类关系抽取算法中,基于词典的方法抽取的关系种类较少,使得概念间的联系相对薄弱,应用于民航应急管理领域时,无法获得大量的概念间的动作关系,很难与实验方法进行对比;基于模式匹配的方法和基于关联规则的方法均不能获知非分类关系的名称,同样无法与实验方法进行对比。因此,本文选择基于模板的方法,在自然语言处理的基础上改进模板的定义方式,定义领域文本集中“主谓宾”形式为非分类关系获取模式,与基于 NNV 关联规则的非分类关系提取方法进行对比,结果如表 3 所示。

表 3 不同方法的非分类关系提取结果对比表

方法	正确非分类关系数 $N_{\text{precision}}$	获取非分类关系数 $N_{\text{algorithm}}$	所有非分类关系数 N_{domain}	准确率 precision/%	召回率 recall/%
模式匹配	31	115	136	27.0	22.8
NNV 关联规则	42	107	136	39.3	30.8

由表 3 可见,基于模式匹配的方法相对简单、过滤条件少,获取的非分类关系数 $N_{\text{algorithm}}$ 较多,但不能保证获取的非分类关系的正确数,导致准确率和召回率都相对较低,而基于 NNV 关联规则的方法得到的准确率和召回率均有所提高。同时,在民航突发事件应急管理领域,应急管理部门重点需要明确各个救援部门的救援动作以指导救援,并且救援动作大都以动词形式存在于实验语料中。因此,利用 NNV 关联规则法,强调动词为对象提取非分类关系的名称,较好地弥补了应急救援时动作提取的空白点,在提高了准确率和召回率的基础上对民航应急管理领域本体的完善提供了良好的支持。

4 结束语

本文在关联规则的基础上,提出了基于 NNV 关联规则的非分类关系获取与实现方法,解决了非分类关系获取中无法自动获取关系名称的问题。由于中文概念的多义性对非分类关系种类的影响等原因,该方法的准确率和召回率在领域本体构建中还可以进一步优化。

参考文献:

[1] 杜小勇,李曼,王珊. 本体学习研究综述[J]. 软件学报,2006,17(9):1837-1847.

[2] 孙明,陈波,周明天. 基于层次关联规则的日志本体事件领域关系学习[J]. 计算机应用研究,2009,26(10):3683-3686.

[3] SANCHEZ D, MORENDO A. Learning non-taxonomic relationships from Web documents for domain ontology construction[J]. Data and Knowledge Engineering,2008,64(3):600-623.

[4] MAEDCHE A, STAAB S. Ontology learning for the semantic Web [J]. IEEE Intelligent Systems,2001,16(2):72-79.

[5] ALFONSECA E, RUIZ-CASADO M, OKUMURA M, *et al*. Towards large-scale non-taxonomic relation extraction: estimating the precision of rote extractors[C]//Proc of the 2nd Workshop on Ontology Learning and Population. Sydney: Association for Computational Linguistics,2006: 49-56.

[6] 杨璇. 民航突发事件应急管理领域本体构建方法的研究与实现[D]. 天津:中国民航大学,2010.

[7] 刘柏嵩. 基于 Web 的通用本体学习研究[D]. 杭州:浙江大学,2007.

[8] 杨丽鹤,林世平. 基于关联规则和自然语言处理技术的概念间非分类关系的抽取[C]//第十三届全国青年通信学术会议论文集. 2008.

[9] 温春,石昭祥,张亮. 中文领域本体概念层次获取方法对比研究[J]. 计算机研究发展, 2009,26(8):2874-2850.