·信息技术与系统·

中文学科术语相关语义关系获取方法研究*

朱 惠[1,2] 杨建林[1,2] 王 昊[1,2]

(1.南京大学信息管理学院 江苏南京 210023)

(2.江苏省数据工程与知识服务重点实验室 江苏南京 210023)

摘 要:文章对中文非结构化文本中(半)自动获取学科术语的相关语义关系进行了探讨,以寻找行之有效的获取方法。从CNKI获取"数字图书馆"学科领域文献,通过术语抽取、动词抽取、向量空间模型构建、双重关联规则分析和规则评价获得了具有较强关联的术语对以及作为关联标签的动词,从而获取了学科术语的相关语义关系。该获取方法与其他方法相比,具有较高的可行性和有效性,并对术语的相关语义关系进行了有效性和实用性的评价,提高了获取的准确率。但文章也存在一定的局限性,在对术语相关语义关系的有效性和实用性进行评价时,指标的选择和阈值的确定存在人工干预,具有一定的主观性。

关键词:学科术语;相关语义关系;数据挖掘;关联规则;规则评价

中图分类号:G202 **文献标识码**:A DOI:10.11968/tsyqb.1003-6938.2017041

Research on the Methods of Extracting Non-taxonomic Relation of Chinese Discipline Terms

Abstract This paper discusses how to (semi-)automatically extract non-taxonomic relation of discipline terms from Chinese unstructured text so as to find feasible and effective extracting methods. First, papers of Digital Library are retrieved from CNKI; then terms and transitive verbs are extracted; third, vector space models are constructed; fourth, association rules are analyzed and evaluated; and last, the term pairs with stronger relation are acquired and the transitive verbs used as the labels of relation, thus the non-taxonomic relation of Chinese discipline terms is extracted. The above method is more feasible and effective than other methods, and it can improve the extracting accuracy by evaluating the effectiveness and practicality. This paper of course has limits. When evaluating the effectiveness and practicality of association rules, the indicators and thresholds are determined by manual intervention, so the method has subjectivity to some extent.

Key words discipline terms; non-taxonomic relation; data mining; association rules; rules evaluating

1 引言

学科术语的语义关系总体上可分为两大类:分类语义关系(层次语义关系)和非分类语义关系,本文将非分类语义关系称为相关语义关系。层次语义关系和相关语义关系均是学科知识本体的重要组成部分,它们将学科术语按照语义关系进行组织,为学科知识的搜索、重用及进一步理解提供条件[1]。在文献[1]中,作者对如何借助知识自动获取方法和技术获得领域术语的层次语义关系进行了研究,本文将探讨如何从中文非结构化文本中(半)自动获得学科术语的相关语义关系。

相较于层次语义关系,相关语义关系的获取更为困难,目前国内外对此的研究也较少,常用的获取相关语义关系的方法之一是普通关联规则分析。该方法能获取术语的相关语义关系,但只能获得具有相关语义关系的术语对,而不能获得关系的标签[2]。

本文将术语的相关语义关系限定为<术语1-动词-术语2>的三元组关系,试图在建立句子-术语向量空间模型和句子-<术语,动词>向量空间模型的基

*本文系江苏省社会科学基金一般项目"领域术语语义关系自动获取研究"(项目编号:15TQB009)与国家自然科学基金青年项目"面向学术资源的TSD与TDC测度及分析研究"(项目编号:71503121)研究成果之一。

收稿日期:2016-10-08;**责任编辑**:魏志鹏

础上,引入双重关联规则分析以及规则评价,由此形成一种从中文非结构化文本获取学科术语相关语义关系的具体方法。双重关联规则分析还没被发现应用在学科术语相关语义关系的获取中,因此,本文尝试引入该方法获得术语的相关语义关系,并借助相关指标来评价规则的有效性和实用性。

2 相关研究

国内外有学者对基于非结构化文本获取术语的相关语义关系进行了研究。如 David 等[3]提出了一个自动的、无监督的获取概念相关语义关系的方法,该方法从网络文本提取术语的相关语义关系,并通过与 Wordnet 进行比较验证方法的有效性;J. Villaverde 等[4]对领域文本语料库进行分析,抽取连接概念对的动词,并将这一技术集成到了本体构建的过程中;Albert 等[5]通过集成类似 DBpedia 这样的外部知识源到本体学习系统中获得相关语义关系的标签。该方法应用了语义推理和验证,使得获取的相关语义关系质量较高;Mei Kuan Wong 等[6]提出基于一种多步骤相关研究框架从非结构化文本中获取术语的相关语义关系;Ivo Serra 等[7]采用两个过程对获取概念相关语义关系的多种技术和方法进行了评价,并在生物学领域语料库和法律领域语料库中进行了验证;Martin 等使用扩展的关联规则获取术语的相关语义关系以及给出了关系的标签,并且基于已有语义标注的语料库对方法进行了评估[8]。

董丽丽等[9]首先通过关联规则抽取特定领域术语对,接着抽取术语对之间的高频动词,将它们作为候选相关语义关系标签,然后运用 VF×ICF 度量方法确定相关语义关系的标签;古凌岚等[10]运用语义角色标注和依存语法分析获取文本句子的语义依存结构,提取出具有语义依存关系的动词框架,通过语义相似度计算发现动词框架中术语间的相关语义关系和关系标签;邱桃荣等[11]通过分析概念粒的上下文,构建了基于不同领域概念粒度空间的概念粒交叉关系学习模型,有利于实现领域本体相关语义关系的获取;王红等[12]提出了基于 NNV(名词-名词-动词)的关联规则获取术语相关语义关系及其标签的方法;张立国等[13]对语料进行词性标注和语义分析,得到具有语义依存关系的动词框架,然后再计算句子的相似度,抽取出术语的相关语义关系并给出关系的标签;谷俊等[14]在关联规则中加入谓语动词进行相关计算,结合搜索引擎技术抽取候选相关语义关系,在此基础上对置信度和支持度进行对比分析,抽取出最终的相关语义关系。

综上可知,国内外学者尝试通过关联规则分析、语义依存分析等来获取术语的相关语义关系,而关联规则分析的应用又较多。作为相关语义关系标签的动词的获取还没有形成有效统一的方法。此外,上述方法对于所获规则的有效性和实用性并没有进行评价。

本文将构建句子×术语向量空间模型、句子××术语,动词>向量空间模型,进行二重关联规则分析以获取具有相关语义关系的术语对以及语义关系的标签。在进行关联规则分析的过程中,引入一系列指标来控制规则的有效性和实用性,从而提高术语相关语义关系获取的质量。

3 学科术语相关语义关系获取方法

本文重点探讨基于双重关联规则分析和规则评价从非结构化文本获取术语相关语义关系的方法和过程,这里的非结构化文本由学科期刊论文的标题、摘要和关键词构成,获取思路和方法(见图 1)。

3.1 术语抽取

科研人员是学科术语动态变化过程的直接参与者和见证者,他们撰写的科研文献记载了学科的动态发展过程,文献中的关键词则是学科研究内容的凝练,因此,可以从科研文献的关键词中抽取学科术语。

但笔者给出的关键词具有较大的随意性、不一致性以及误差性,因此,有必要首先对这些候选术语进行统一规范,以符合同一概念的术语唯一化。

学科术语是专业词汇,必须具有一定的学科认可度,因此,本文采用关键词在所有文档中出现的频数 N_k 作为筛选条件,即若:

$$N_k \geq C \quad (1)$$

则认为该关键词被学科普遍认可,可作为该学科的术语,其中 C 为词频阈值[1]。

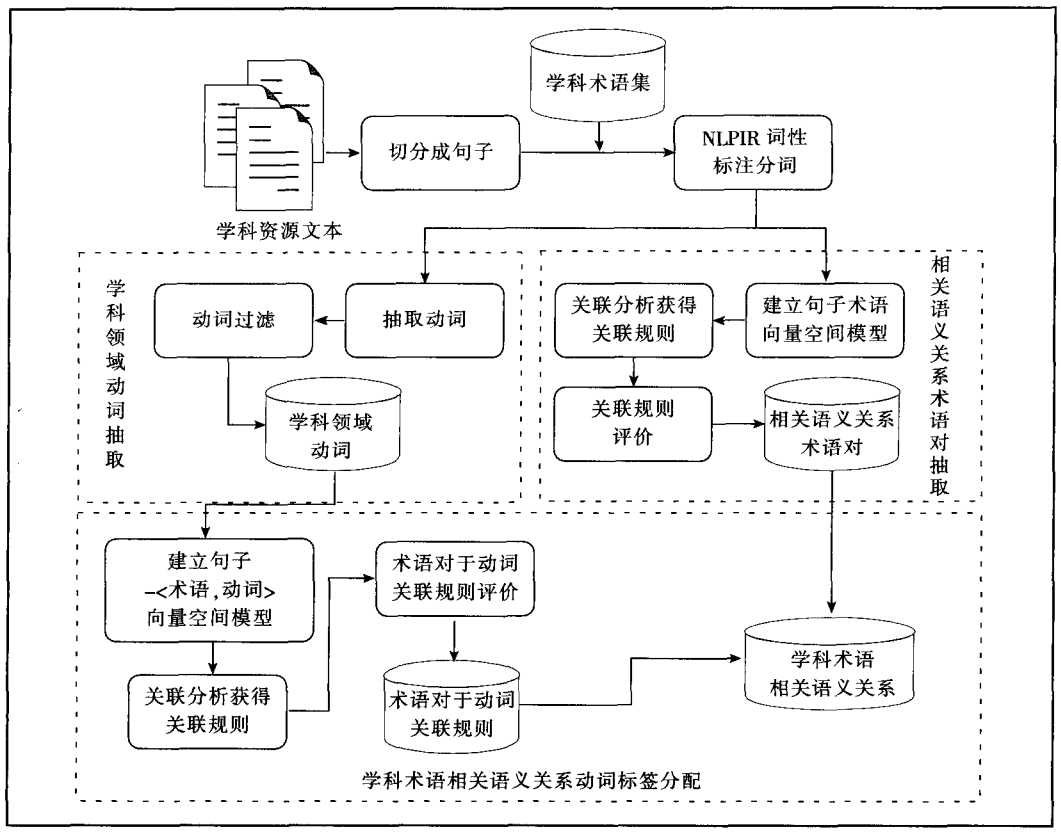

图1 学科术语相关语义关系获取思路和方法

3.2 相关语义关系术语对抽取

以非结构化文本中的摘要作为研究语料，将其切分成句子，进行词性标注分词，构建句子×术语向量空间模型。运用关联规则分析获取具有相关语义关系的学科术语对。在获取过程中，要注意规则的有效性和实用性，本文将借助规则置信度、规则支持度、规则提升度、置信率、正态卡方、信息差这些指标对获得的规则进行有效性和实用性的评价[15]244-249。

把所有句子看成是事务集，而每个句子中包含的术语则是该事务所包含的项目。如果术语1(X)与术语2(Y)在整个事务集中具有一定的共现次数，且术语2在术语1出现的情况下出现了较多次，则认为这样的两个术语具有有效的关联关系。这里引入评价关联规则有效性的指标：规则置信度 $C_{X\to Y}$ 和规则支持度 $S_{X\to Y}$。

进一步，针对具有有效关联关系的术语1与术语2：

(1)如果术语2在术语1中出现的密集程度比它在整个事务集中出现的密集程度还要大，则认为术语1与术语2间的关联关系不仅有效，而且具有实际意义，即具备实用性。这里引入评价关联规则实用性的指标：规则提升度(Lift)和置信率(R)。当Lift大于1时，Lift越大，关联越强。R越大，关联越强。

(2)构建术语1和术语2的二维随机变量(X，Y)，从而计算X与Y之间的统计相关性，如果统计相关性高于某个数值，则认为术语1与术语2间的关联关系不仅有效，而且具备实用性。这里引入评价关联规则实用性的指标：正态卡方(N)。N越接近1，关联越强，N越接近0，关联越弱。

(3)对于上述(2)中的二维随机变量(X，Y)，计算X与Y独立情况下(X，Y)概率分布的信息熵Ent1、(X，Y)实际概率分布的信息熵Ent2，求差E = Ent1-Ent2，如果E大于某个数值，则认为术语1与术语2间的关联关系不仅有效，而且具备实用性。这里引入评价关联规则实用性的指标：信息差(E)。E越大，关联越强。

3.3 动词抽取

在自然语言处理领域，一般认为，动词是句子中最能表达关系的信息。基于这样的认识，动词可以用

来作为标记同一句子中术语间相关语义关系的标签，形成<术语1-动词-术语2>这样的三元组关系，如<数字图书馆-构建-数字空间>、<专家系统-组织-知识>。利用NLPIR中文分词工具对句子语料进行词性标注分词,抽取其中的动词;然后对这些动词进行停用词处理、单字词汇处理以及基于VF-ICF(Verb Frequency-Inverse Concept Frequency)指标的筛选,从而获得学科动词。

VF-ICF是类似于TF-IDF的概念，它的作用也与TF-IDF类似[16]。TF-IDF主要用来度量词汇在文档集中的重要性。VF-ICF则主要用来度量动词在概念对或术语对中的重要性,那些出现频数高且与更多术语对共现的动词,它的重要性被拉低;而那些出现频数低但仅与少数动词共现的动词,它的重要性被抬高。

假设vf_j表示动词v_j在句子集中出现的频数，C表示句子集中术语对的总数目,c_j表示在整个句子集中与动词v_j共现的术语对数目,则动词v_j的权重为:

$$w_j=vf_j\times\log(\frac{C}{c_j}) \quad (2)$$

3.4 相关语义关系标签分配

以3.2中获得的具有相关语义关系的学科术语以及3.3中获得的学科动词作为句子的特征项,构建句子×<术语,动词>向量空间模型,再次运用关联规则分析,以术语对为规则的前项、动词为规则的后项获取术语对与动词的关联规则。为保证术语对内部有较强关联，应按以下规则筛选：剔除那些关联规则,这些规则的术语对没有出现在3.2的779术语对中。进一步借助规则置信度等相关指标对获得的关联规则的有效性和实用性进行评价，最终获得这样的一些关联规则：术语对与学科动词具有较强关联关系,同时,术语对中的两个术语也具有较强关联关系。这样便获取了具有相关语义关系的术语对及其动词标签。

4　实验结果及分析

本文以“数字图书馆”学科领域的期刊论文作为分析对象，基于<句子-术语>语义关联以及<句子-术语,动词>语义关联进行双重关联规则分析,并在分析过程中引入相关指标来评价规则的有效性和实用性。

4.1 数据预处理

以“数字图书馆”为主题词,在CNKI中国期刊全文数据库的核心期刊范围内检索1996至2011这15年间发表的论文,共计6446篇。抽取标题、摘要和关键词构成非结构化文本。通过术语抽取最终获得911个术语[1]。

从6446篇非结构化文档中提取摘要部分,将其切分为28094个句子,剔除长度小于6的那些句子,共获得27056个句子。以学科术语集为用户词典,对27056个句子利用NLPIR中文分词工具进行分词,共获得61114个句子术语对。那些只含有1个术语的句子,不能从中抽取出相关语义关系,因此,剔除掉这些句子,共获得16608个句子,涉及术语911个。

以这911个学科术语为用户词典，利用NLPIR对16608个句子进行词性标注分词，共得到47060个动词词汇。这些动词包括及物动词v、名动词vn、副动词vd、趋向动词vf、动词性语素vg、不及物动词vi、动词性惯用语vl、是动词vshi、有动词vyou和形式动词vx。

由于用作相关语义关系标签的动词必须连接两个术语,因此,本文选择及物动词v作为候选学科动词,共1312个,对它们进行进一步筛选：

(1)去除停用词。1312个动词词汇去除停用词后还剩下1249个词汇。

(2)去掉长度为1的单字动词词汇。笔者经过对单字动词词汇的观察，认为这样的动词并不能很好地表达术语间的相关语义关系，因此剔除掉这些词汇,还剩下1059个词汇。

(3)选择在整个句子集中出现一定频数以上的那些动词词汇。在1059个词汇中,有368个词汇仅出现了1次,笔者认为这些低频数出现词汇的代表性较差,需要剔除,最终获得了691个候选学科动词。

4.2 第一重关联规则分析

以16608个句子和911个术语构建了16608行×911列的句子术语向量空间模型。采用数据挖掘工具Clementine,基于Apriori算法进行关联规则分析。

关联规则分析是一种无监督的学习方法，评价

规则有效性和实用性的指标阈值的设置均要依靠领域专家的专业知识并结合所分析的实际问题来确定。笔者在进行关联规则分析时,对各指标阈值的取值进行了相关的尝试。

4.2.1 有效规则筛选

表1列出了不同规则置信度和不同规则支持度下的关联分析结果,置信度和支持度交叉位置单元格内的数值是在相应条件下抽取到的关联规则数目。

经过对不同规则置信度和规则支持度下结果的观察,结合领域专家的意见,并考虑置信度和支持度的取值,笔者最终选定了规则置信度≥30%且规则支持度≥0.01%取值条件下的分析结果,共得到971条有效的关联规则,这些规则共涉及术语658个。

4.2.2 实用规则筛选

(1)在获得的971条有效关联规则中,规则提升度的最小值是1.11,最大值是4152.00,平均值是125.53。最小值是1.11表明所有的规则提升度均大于1,说明后项在前项中出现的概率大于后项在整个事务集中出现的概率,这样的规则有一定的实际意义(所有规则提升度取值情况见表2)。

由表2数据可知,规则提升度的取值范围很广,说明规则置信度与后项支持度取值的差异性较大,这是由数据的稀疏性导致的。在本文的数据中,有些后项Y在整个事务集中覆盖的范围很窄,出现的频数很低,这就导致了这些后项的支持度取值较低,进一步导致规则提升度很高。规则提升度取值范围太大会给筛选规则带来困扰,而且不同的样本数据会有不同的取值范围。为了克服这个问题,可以对规则提升度标准化:置信率把规则提升度压缩在[0,1)区间内。

(2)置信率是由规则提升度转变而来,它更适合于对稀疏样本的分析。笔者对971条关联规则的置信率进行了计算,最小值为0.0991,最大值为0.9998(971条有效关联规则置信率取值的频数分布见表3)。

领域专家在设置置信率阈值的时候,可以根据实际问题的具体情况确定,在本文的分析中,笔者将置信率的阈值设置为0.5,即选取置信率大于等于0.5的那些关联规则。经筛选后,共得到779条关联规则,涉及术语568个。

(3)对经过置信率筛选后得到的779条关联规则进行正态卡方的计算,其中最大值为1,最小值为0.0001。大部分的正态卡方取值较小(正态卡方取值的频数分布情况见表4)。

由表4可知,有8个关联规则的正态卡方值为1,其中包括"社会阅读"→"图书馆法治"(0.06%,100.00%)和"图书馆法治"→"社会阅读"(0.06%,100.00%)。这两条规则的置信度均为100.00%,取值

表1 基于Apriori算法不同规则置信度和规则支持度下关联分析结果

规则数目		规则支持度									
		≥0%	≥0.01%	≥0.02%	≥0.03%	≥0.04%	≥0.05%	≥0.06%	≥0.07%	≥0.08%	≥0.09%
规则置信度	≥20%	3079	1879	987	772	536	414	385	259	213	194
	≥30%	1441	971	582	468	322	247	228	122	94	85
	≥40%	840	599	377	307	220	174	163	68	48	42
	≥50%	609	368	242	201	147	119	114	35	23	19
	≥60%	273	220	152	137	101	83	83	17	10	7
	≥70%	189	136	107	102	79	69	69	12	6	4
	≥80%	143	90	70	65	54	49	49	7	3	2
	≥90%	120	67	47	45	37	34	34	3	2	1

表2 规则提升度取值频数分布

规则提升度	规则数	频率(%)	累积规则数	累积频率(%)	规则提升度	规则数	频率(%)	累积规则数	累积频率(%)
(1000, 4152)	39	4.02	39	4.02	(20, 50]	126	12.98	370	38.11
(500, 1000]	48	4.94	87	8.96	(10, 20]	70	7.21	440	45.31
(200, 500]	37	3.81	124	12.77	(5, 10]	139	14.32	579	59.63
(100, 200]	53	5.46	177	18.23	(2, 5]	200	20.60	779	80.23
(50, 100]	67	6.90	244	25.13	(1, 2]	192	19.77	971	100.00

相同。前条规则的100.00%置信度说明“社会阅读”出现的时候必出现“图书馆法治”,同理,后条规则的100.00%置信度说明“图书馆法治”出现的时候也必出现“社会阅读”,因此,这两个术语在文档中要么不出现,要么一起出现,它们具有最强的关联关系。其他7条规则也是类似的情况。规则支持度为0.06%,说明术语“社会阅读”和“图书馆法治”在整个事务集(16608个事务)中共现了10次。

笔者也对所有规则的正态卡方与提升度、置信率间的关系进行了考察,结果表明,正态卡方与规则提升度和规则置信率并不冲突,可以依据其取值的排序来评价关联规则关联关系的强弱。

(4)笔者计算了所有779条关联规则的信息差,其中最大值为0.01283,最小值为0.00007(所有信息差取值的频数分布见表5)。

对信息差和正态卡方这两个评价指标进行相关性分析,结果表明,这两个指标具有统计学意义上的显著相关性。因此,在评价关联规则实用性的时候,可以综合规则提升度、规则置信率、正态卡方和信息差这些指标对规则进行筛选。

最终,笔者结合以上4个评价关联规则实用性指标的取值以及对具体关联规则的实际观察,共抽取出了779个具有关联关系的术语对,涉及术语568个(部分术语对见表6)。

4.3　学科动词筛选

依据公式2可计算出所有候选学科动词的权重,领域专家可以根据实际情况确定阈值W,选取wj大于等于W的那些动词作为学科动词。笔者根据公式2对691个候选动词进行权重计算(部分计算结果见表7)。

笔者根据实际情况选取W=20,剔除了128个动词,最终获得563个学科动词。

4.4　第二重关联规则分析

基于16608个句子、911个学科术语以及563个学科动词建立了16608行×1474列的向量空间模型。运用Apriori算法进行关联分析的时候,以术语对为前项,动词为后项,规则置信度$C_{X\to Y}$阈值设定为10%,规则支持度$S_{X\to Y}$阈值设定为0.01%(保证术语对与动词在整个句子集中至少共现2次),共获得了43913个关联规则。

在这些关联规则中,有些前项中的两个术语之间并没有较强的关联关系,因此,须对这些关联规则

表3　关联规则置信率取值频数分布

置信率	规则数	频率(%)	累积规则数	累积频率(%)	置信率	规则数	频率(%)	累积规则数	累积频率(%)
[0.9,1)	440	45.31	440	45.31	[0.4,0.5)	42	4.33	821	84.55
[0.8,0.9)	139	14.32	579	59.63	[0.3,0.4)	57	5.87	878	90.42
[0.7,0.8)	102	10.50	681	70.13	[0.2,0.3)	37	3.81	915	94.23
[0.6,0.7)	64	6.59	745	76.73	[0.1,0.2)	52	5.36	967	99.59
[0.5,0.6)	34	3.50	779	80.23	[0,0.1)	4	0.41	971	100.00

表4　关联规则正态卡方取值频数分布

正态卡方	规则数	频率(%)	累积规则数	累积频率(%)	正态卡方	规则数	频率(%)	累积规则数	累积频率(%)
1	8	1.03	8	1.03	[0.005, 0.01)	75	9.63	336	43.13
[0.5, 1)	22	2.82	30	3.85	[0.001, 0.005)	249	31.96	585	75.10
[0.1, 0.5)	66	8.47	96	12.32	[0.0005, 0.001)	98	12.58	683	87.68
[0.05, 0.1)	39	5.01	135	17.33	[0.0001, 0.0005)	96	12.32	779	100.00
[0.01, 0.05)	126	16.17	261	33.50					

表5　关联规则信息差值的频数分布

信息差	规则数	频率(%)	累积规则数	累积频率(%)	信息差	规则数	频率(%)	累积规则数	累积频率(%)
[0.01, 0.02)	4	0.51	4	0.51	[0.0005, 0.001)	224	28.75	487	62.52
[0.005, 0.01)	66	8.47	70	8.99	[0.0002, 0.0005)	214	27.47	701	89.99
[0.002, 0.005)	65	8.34	135	17.33	[0, 0.0002)	78	10.01	779	100.00
[0.001, 0.002)	128	16.43	263	33.76					

进行过滤:前项中的两个术语必须是4.2中获得的术语对。经过滤后,共获得779条关联规则。

再次利用规则提升度对关联规则进行筛选:值大于等于2,经筛选后,共获得770条关联规则。因此,这些关联规则反映了术语的相关语义关系的术语对及其动词标签(部分结果见表8)。

因为评价关联规则有效性和实用性时对相关指标阈值的设定完全由领域专家人为决定,因此具有一定的主观性。领域专家应充分了解学科术语特点以及数据的特征,进行合理的设定。

4.5 与其他方法及技术比较

目前,从领域非结构化文本中抽取领域术语相关语义关系的研究较少,采取的其他方法一般有:(1)基于词汇-句法模式的方法;(2)基于句法分析的方法。第(1)种方法必须人工制定获取模板,因此获得的相关语义关系受制于模板的准确性和完备性;第(2)种方法要求对句法进行分析,由于中文语法句法的复杂性,实现较为困难。

本文所采用的二重关联规则分析结合规则评价的方法具有较高的可行性和有效性,不仅能从非结构化文本中获取学科术语的相关语义关系及其标签,还能评价语义关系的有效性和实用性。

5 结语

本文提出了一种从学科非结构化文本获取学科术语相关语义关系的方法,该方法通过术语抽取、动词抽取、向量空间模型构建、二重关联规则分析和规则评价获取术语的相关语义关系及其标签。该方法基于句子-<术语,动词>向量空间模型运用关联规则分析获取相关语义关系的标签,并借助规则支持度、规则置信度、置信率等指标对关联规则的有效性和实用性进行控制。本文所采用的方法与其他方法相比具有以下明显优势:能更行有效地获得相关语义关系的标签,并对相关语义关系的质量进行控制。本文以"数字图书馆"学科领域为例论证了该方法的可行性和有效性,但也存在一些缺陷,评价指标的选择

表6 基于Apriori算法关联分析抽取出的具有相关语义关系的术语对片段

术语1	术语2	术语1	术语2	术语1	术语2
社会阅读	图书馆法治	智能图书馆	主动信息服务	电子图书	内容格式
数字典藏	引文分析	风险管理	风险识别	知识发现	信息发现
动态定制	网络服务组合	书目数据库	古籍数字化	档案馆	非物质文化遗产
社会阅读	知识自由	信息用户	客户关系管理	全文检索	系统架构
语义网服务	OWL-S	知识社区	后数字图书馆	检索技术	系统集成
FCSAN	IPSAN	分类法	主题词表	古籍	书目数据库
信息公平	知识自由	3G	移动服务	推荐系统	协同过滤
网络信息组织	自主型学习	流媒体	复合数字对象	信息安全	风险评估
文献计量	内容分析	运行机制	知识市场	数据库建设	网络平台
智能图书馆	普适计算	图书情报学	引文分析法	数据挖掘	关联规则

表7 候选学科动词权重wj取值片段

动词vj	频数vfj	术语对数目Cj	权重wj	动词vj	频数vfj	术语对数目Cj	权重wj
筹措	2	81	5.67	建立	449	3211	554.42
净化	2	69	5.81	利用	451	2664	593.46
趋同	2	69	5.81	论述	477	2401	649.21
报道	2	67	5.83	阐述	450	1879	660.37
封闭	2	67	5.83	提供	572	3611	677.13
抓取	2	49	6.10	实现	714	3810	828.59
做到	2	46	6.16	探讨	689	3201	851.70
简化	2	38	6.32	介绍	806	3584	956.76
带动	2	37	6.35	分析	850	3592	1008.17
感到	2	37	6.35	提出	1395	6640	1282.36

表 8　具有相关语义关系的术语对及与其动词标签

术语对编号	术语 1	动词	术语 2	置信度 $C_{X\to Y}$ (%)	支持度 $S_{X\to Y}$ (%)	提升度 Lift
1	数字图书馆	提出	数字空间	40.000	0.012	4.776
	数字图书馆	构建	数字空间	40.000	0.012	18.052
2	XML	实现	RDF	14.286	0.012	3.385
	XML	解决	RDF	14.286	0.012	7.962
	XML	描述	RDF	14.286	0.012	23.965
3	OAI-PMH	提出	互操作	28.571	0.012	3.411
	OAI-PMH	实现	互操作	42.857	0.018	10.154
4	数字图书馆	提出	代理技术	40.000	0.012	4.776
	数字图书馆	应用	代理技术	40.000	0.012	46.133
5	图书馆	探讨	虚拟馆藏	28.571	0.012	6.887
	图书馆	处理	虚拟馆藏	28.571	0.012	105.448
	图书馆	拥有	虚拟馆藏	28.571	0.012	279.126
	图书馆	存取	虚拟馆藏	28.571	0.012	474.514
6	用户建模	提出	个性化服务	66.667	0.012	7.960
	用户建模	设计	个性化服务	66.667	0.012	147.627
7	知识构建	提出	知识服务	40.000	0.012	4.776
	知识构建	实现	知识服务	40.000	0.012	9.477
8	专家系统	提出	知识	50.000	0.012	5.970
	专家系统	采用	知识	75.000	0.018	73.704
	专家系统	组织	知识	75.000	0.018	197.714
9	图书馆	使用	版权法	40.000	0.012	28.512
10	数字图书馆	提供	个性化服务	14.286	0.066	4.244

和阈值的确定存在人工干预，带有一定的主观性。在今后的研究工作中，笔者将进一步尝试运用不同的机器学习方法(半)自动获取学科术语的相关语义关系,探讨更有效可行的策略和方案。

参考文献:

[1] 朱惠,杨建林,王昊.中文领域专业术语层次关系构建研究[J].现代图书情报技术,2016(1):73-80.

[2] Maedche A,Staab S.Discovering Conceptual Relations from Text[A].Proc.of the 12th International Conference on Software and Knowledge Engineering[C].Berlin,Germany:[s.n.],2000:321-325.

[3] David Sa'nchez,Antonio Moreno.Learning non-taxonomic relationships from web documents for domain ontology construction[J].Data & Knowledge Engineering,2008,64(3):600-623.

[4] J.Villaverde,A.Persson,D.Godoy,etal.Supporting the discovery and labeling of non-taxonomic relationships in ontology learning[J].Expert Systems with Applications,2009,36(7):10288-10294.

[5] Albert Weichselbraun,Gerhard Wohlgenannt,Arno Scharl.Refining non-taxonomic relation labels with external structured data to support ontology learning[J].Data & Knowledge Engineering,2010,69(8):763-778.

[6] Mei Kuan Wong,Syed Sibte Raza Abidi,Ian D.Jonsen.A multi-phase correlation search framework for mining non-taxonomic relations from unstructured text[J].Knowledge and Information Systems,2014,38(3):641-667.

[7] Ivo Serra,Rosario Girardi,Paulo Novais.Evaluating techniques for learning non-taxonomic relationships of ontologies from text[J].Experts Systems With Applications,2014,41(11):5201-5211.

[8] Martin KAVALEC,Vojtech SVATEK.A Study on Automated Relation Labelling in Ontology Learning[EB/OL].[2016-10-15].http://nb.vse.cz/~svatek/olp05.pdf.

(上转第 5 页)

式展示了自古以来20位中华名人的家训和家风故事，吸引到馆读者们驻足凝思。3月下旬，借全国多位知名专家学者齐聚深圳审议确定2017年“南书房家庭经典阅读书目”之机，深圳图书馆又联合深圳实验教育集团、中学生文联，邀请部分与会专家走进校园，面向高中师生举办了一场以“为什么读经典”为主题的别开生面的高端对话会。深圳图书馆官方微博对这场活动首次尝试现场直播，短短两小时的活动使现场与直播观众四千余人经历了生动精彩、谈经论典的“经典”时刻，这样的活动接地气、有生气，值得图书馆人以“走出去”的精神不断探索。

围绕“家庭”“经典”“阅读”等关键词，图书馆与家庭阅读、家庭教育已产生广泛关联，在资源保障、人才培养、空间利用、平台搭建等诸方面更积极联动，产生良好的推广效应。汲取传统家庭教育的智慧，施以图书馆这座资源宝库与人才队伍的援手，相信家庭阅读之风会得到越来越多的关注，家庭文化传承后继有人。

参考文献：

[1] 马建光.钱氏家族英才辈出的文化密码[J].领导科学，2016(28)：46-48.
[2] 刘晓飞，廉武辉，刘小艳.从“家风”建设看梁启超的“梁氏家教”[J].教育文化论坛，2016(2)：8-12.
[3] 刘胜梅.曾氏家风的内涵及其现实启示[J].学术探索，2016(4)：127-132.
[4] 胡晓.李鸿章家族文化述论[J].合肥教育学院学报，2000(1)：33-37.
[5] 凤凰村古书室书写“文”脉传奇[N/OL].[2017-03-01].http://www.sznews.com/zhuanti/content/2014-04/22/content_9395700.htm.
[6] 刘丽川.深圳客家研究[M].深圳：海天出版社，2013:158
[7] 蒋荣耀.坑梓客家围亟待抢救保护[N].深圳商报，2017-03-07(10).
[8] 林语堂.苏东坡传[M].天津：百花文艺出版社，2000：29-30.
[9] 包世臣.包世臣全集·安吴四种总目序[M].合肥：黄山书社，1993.
[10] 王爽，宫丽颖.数字时代下影响儿童阅读的因素分析：深度解析美国《2014儿童与家庭阅读报告》[J].出版参考，2015(11)：24-26.

作者简介：张岩，女，深圳图书馆研究馆员。

（下接第132页）

[9] 董丽丽，胡云飞，张翔.一种领域概念非分类关系的获取方法[J].计算机工程与应用，2013，49(4):157-161.
[10] 古凌岚，孙素云.基于语义依存的中文本体非分类关系抽取方法[J].计算机工程与设计，2012，33(4):1676-1680.
[11] 邱桃荣，黄海泉，段文影，等.非分类关系学习的粒计算模型研究[J].南昌大学学报(工科版)，2012，34(3):273-278.
[12] 王红，高斯婷，潘振杰，等.基于NNV关联规则的非分类关系提取方法及其应用研究[J].计算机应用研究，2012，29(10):3665-3668.
[13] 张立国，陈荔.维基百科中基于语义依存的领域本体非分类关系获取方法研究[J].情报科学，2014，32(6):93-97.
[14] 谷俊，严明，王昊.基于改进关联规则的本体关系获取研究[J].情报理论与实践，2011，34(12):121-125.
[15] 薛薇，陈欢歌.Clementine数据挖掘方法与应用[M].北京：电子工业出版社，2010:244-249.
[16] 舒万里.中文领域本体学习中概念和关系抽取的研究[D].重庆：重庆大学，2012.

作者简介：朱惠(1979-)，女，南京大学信息管理学院讲师，博士，研究方向：信息智能处理与检索、知识本体构建及应用、数据挖掘；杨建林(1970-)，男，南京大学信息管理学院教授，研究方向：信息智能处理与检索、信息分析评价、数据挖掘；王昊(1981-)，男，南京大学信息管理学院教授，研究方向：信息智能处理与检索、知识本体构建及应用、科学评价和引文分析。