情 报 学 报 ISSN 1000 - 0135
第 32 卷 第 7 期 708 - 715, 2013 年 7 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 32 No. 7, 708 - 715 July 2013

doi:10.3772/j.issn.1000-0135.2013.07.004

基于词形规则模板的术语层次关系抽取方法[1)]

韩红旗[1] 徐 硕[1] 桂 婕[1] 乔晓东[1] 朱礼军[1] 安小米[2]

(1. 中国科学技术信息研究所,北京 100038;
2. 数据工程与知识工程教育部重点实验室(中国人民大学),中国人民大学信息资源管理学院,北京 100872)

摘要 术语层次关系抽取是领域概念关系体系构建的重要基础。针对目前术语关系抽取中手工实现的问题,提出了基于词形规则模板匹配的术语层次关系抽取方法,实现从科技论文文本中抽取类属关系(IS-A)和整体部分关系(PART-OF)关系。利用复合术语的 head 和 modifier 特征,比较两个术语之间存在的边缘共用词汇,构造模板来确定它们之间的 IS-A 和 PART-OF 关系;提出泛化度指标,用于测量两个术语在概念层次树上的相对位置;提出相关度概念,用于测量两个术语之间在语义上的相关性。对不存在共用词汇和不匹配模板的术语采用泛化度差值和相关度来判断它们之间是否存在层次关系。实验从信息资源管理领域的论文文本中提取层次关系术语对 1306 对,准确率达到 92.5%,证明提出的方法是有效的。

关键词 术语关系抽取 层次关系 词形规则 文本挖掘

Term Hierarchical Relation Extraction Method Based on Morphology Rule Template

Han Hongqi[1], Xu Shuo[1], Gui Jie[1], Qiao Xiaodong[1], Zhu Lijun[1] and An Xiaomi[2]

(1. *Institute of Scientific and Technical Information of China (ISTIC), Beijing* 100038, *China*;
2. *Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE; School of Information Resource Management, Renmin University of China Beijing* 100872, *China*)

Abstract A term relationship extraction method was put forward integrating morphology rules and statistics analysis to extract two types of hierachical relations among term pairs: IS-A and PART-OF. In morphology rule analysis, five types of templates were designed to judge hierachical relations among terms with common left or right sub-string using head or modifier feature of multi terms. A generation index was put forward to measure generation degree of a term and judge hierachical position of two terms; an association index was put forward to measure association degree of term pair and judge similarity relation of two terms in concept tree. Presented methods contains following process: term generation measure computation, term pair association degree computation, morphology rule analysis and template match, non-match term pair relationship judgement. In experiments, 1306 hierarchical relation term pairs were extracted from information resource management paper corpus, and the precision is 92.5%.

Keywords term relation extraction, hierarchical relation, morphology rule, text mining

---

收稿日期:2012 年 9 月 17 日

作者简介:**韩红旗**,男,1971 年,博士,讲师,主要研究方向:文本挖掘、科技情报分析。E-mail:bithhq@163.com。**徐硕**,男,1979 年生,博士,副研究员,主要研究方向:文本挖掘,信息抽取;**桂婕**,女,1976 年生,博士,副研究员,主要研究方向:数据挖掘,专利分析;**乔晓东**,男,1966 年生,英国谢菲尔博士学硕士,研究员,主要研究方向:信息服务,信息资源管理等;**朱礼军**,男,1973 年生,博士,研究员,主要研究方向:知识组织、语义检索等;**安小米**,女,1965 年生,博士生导师,教授,主要研究方向:知识管理、信息资源管理。

1) 本研究受中国科学技术信息研究所预研项目"基于内容和链接的学术社交网络分析"(YY201221)、"十二五"国家科技支撑计划项目"面向外文科技知识组织体系的大规模语义计算关键技术研究"(2011BAH10804)"基于 STKOS 的知识服务应用示范"(2011BAH10B06)、中国人民大学明德学者科学研究基金(中央高校基本科研业务费专项资金资助)"知识工程背景下信息资源管理术语构建方法研究"项目(10XNJ052)资助。

1 引 言

近年来,本体学习技术成为计算机科学领域和知识组织领域的一个研究热点。术语和术语关系的抽取是本体构建的基础,现在大多归于本体构建技术类[1]。术语关系抽取是指从一定规模的语料中抽取出能反映某一领域文本特征的两两词语间的语义关系(同义关系、上下位关系、整体 - 部分关系等)[2]。术语语义关系集中体现和负载了一个学科领域的核心知识,对了解和把握一个学科领域的现状、发展趋势等具有重要的理论和实际意义[3]。术语关系可以应用到机器翻译、信息检索、本体构建等领域,为面向领域的智能知识系统提供支持。在一个本体概念系统模型中,最重要的两种关系是层次关系(hierarchical relations)和关联关系(associative relations)[4],关联关系是因为时间或空间相似,可以凭借经验在概念之间建立专题的连接,而层次关系包含类属关系(generic relations 或 IS-A)和整体部分关系(partitive relations 或 PART-OF)两种,是概念本体中最基础的关系。

概念之间的层次关系是本体中一个重要的组成部分,但目前主要是采用手工实现[5],因此非常有必要研究半自动化或自动化的概念关系抽取方法,来解决手工方法工作量大、周期长等问题。非手工方法获取词汇语义关系主要有两种方法:分布的概率统计方法[6~8]和基于模式方法[9~11]。统计方法利用词语上下文信息,根据一些经典的统计分布假设,计算词语间相关性。这种方法从某种程度上表达了词语之间一种宽泛的关系,它不能精确地定义词语之间具体的语义关系,只是提供二者具有某种关系的佐证。基于模式的方法通过发现词语同时出现的固定模式,用这种模式来直接地表示某种固定的语义关系。实际实验中,模式匹配的方法找到的上下位关系和整体部分关系等的正确率更高,但是模式在文本中出现的频率较低,因此需要更多包含目标词语对的句子,以找到包含此种模式的实例。而统计的方法能够对语料中包含目标词语对的句子极尽其用,因此同样适合于非大规模语料。从总体上看,基于模式的方法和基于统计的方法在抽取语义关系时能够很好的互补[12]。

Hearst[11]较早对术语关系抽取方法进行了研究,她通过研究发现,一些关键的词汇能够很好地指示出上下位关系(hyponym/hypernym)。Grefenstette[13]的研究发现,通过使用一些简单的语法分析,可以将一些名词和动词组合的词语划分到 9 个类别中;Woods[14]也进行了类似的研究,他发现能够根据名词的头部(head)或修饰语(modifier)来决定所属的类别。受 Woods 研究的启发,张巍[15]按照两个术语在词形上的相同之处,定义了一组较通用的特征模板(见表 1)来判定两个术语之间的层次关系且取得了较好的效果。

在张巍的研究中,认为符合 T2 和 T4 模板的两个术语之间一定存在层次关系,而本研究不认同这种观点。例如,“医学信息资源管理”和“图书馆人力资源管理”,“human resource management”和“water resources management”均满足 T2 模板,但它们显然不存在 IS-A 关系;又如,“信息资源中心”和“信息资源管理政策法规”满足 T4 模板,但显然不存在 PART-OF 关系。所以,对于满足 T2 或 T4 模板的两个术语,仍需判定其是否存在 IS-A 或 PART-OF 关系。另一个方面,这些模板不能解决不符合上述模板的层次关系的抽取,也不能确定两个术语哪个代表了上位概念、哪个代表了下位概念。

表 1 术语层次关系模板

模板编号	模板	模板类型	模板实例
T1	(C, A + C)	IS-A	(情报领域, 图书情报领域)
T2	(A + C, B + C)	IS-A	(信息技术, 数据挖掘技术) (人力资源开发, 数字信息资源开发)
T3	(C + B, C)	PART-OF	(多媒体信息资源管理, 多媒体信息资源)
T4	(C+ A, C + B)	PART-OF	(信息资源管理, 信息可视化) (信息资源管理教育, 信息资源管理开发)
T5	(A+ B+ C, A+ C)	IS-A	(信息资源管理技术,信息管理技术) (现代信息资源管理,现代企业信息资源管理)

Pum-Mo Ryu[16]的研究指出术语的专用性(specificity)是术语包含特定领域信息的数量,特定领域术语较一般术语包含更多的领域信息,这种特定信息是构造领域概念层级树的重要条件。陈珂[17]的研究认为在一个层级关系的概念体系中,两个具有直接上下位关系的概念具有较高的语义相似度,一般来说他们在语料库中存在较高的重现率,可以用来评价两个术语在概念层次树上的相对位置。受到Pum-Mo Ryu和陈珂研究的启发,本研究提出了泛化度指标和相关度指标,来解决张巍提出的方法[15]中的问题。泛化度指标用于测量两个术语在概念层次树上的相对位置,相关度用于测量两个术语之间在语义上的相关性。

2 术语的泛化度和相关度

术语关系抽取的理论依据是:如果两个概念在语料库中所处的上下文语言环境总是非常相似,那么可以认为它们之间存在着某种语义关系[18],因此术语层次关系抽取的核心在于确定两个术语之间是否存在概念意义上的层次关系以及相关关系。只有存在层次关系且概念相关的两个术语之间才可能存在语义层级关系。提出了泛化度指标测量来判定术语之间的语义层次性,提出了相关度指标来判定术语之间的语义相关性。

2.1 术语的泛化度

概念层级树上的上下位关系是一种泛化关系,父结点和子结点相比是更抽象的概念,子结点和父结点相比则是更具体的概念。这种概念之间的泛化关系是一种偏序关系,如果用 $c_1 > c_2$ 表示概念 c_1 比 c_2 更泛化,用 $c_2 > c_3$ 表示概念 c_2 比 c_3 更泛化,则存在 $c_1 > c_3$,即概念 c_1 比 c_3 更泛化。将概念的这种层次关系引申到术语之间的层次关系上。如果一个术语 t_1 对应的概念 c_1 比另一个术语 t_2 对应的概念 c_2 更泛化,则说明在概念层级树上 c_1 比 c_2 更靠上,也就是说深度更低一些,反之则说明相反的情况[16]。假设在一个领域的语言系统中,这种语义关系存在一致性,也就是说,即便两个概念意义相距较远、不是很近的上下层关系(如图1中的 t_1 和 t_3),也可以从层级上分辨出两个概念的泛化度,那么就可以采用泛化度来测量任意两个术语之间的层级关系。

一般来说,一个术语所代表的概念越抽象,越会在研究领域的多个子领域出现和使用,但在某一个特定领域中的专用性就较低,反之,如果代表术语的概念越具体,它往往只在特定领域或较少子领域出现和使用,在特定领域的专用性就越高。例如,管理是一个具有较高泛化度的概念,在很多领域通用,但信息资源管理则是一个泛化度较低的概念,在信息资源领域具有专用性。基于上述的术语在研究领域中的这种使用现象,引入术语的泛化度概念,把一个术语在一个领域中的专用性称为术语的泛化度(Generalization)。在一个语料库中,如果把一个术语 t_i 的出现作为输出通道上观察到的一个消息,即把 t_i 的出现作为一个事件,那么就可以把 t_i 出现这一事件所具备的信息量 $H(t_i)$ 作为测量其专用性的测量[16]。借鉴信息论中信息量的定义公式,定义一个术语 t 的泛化度公式如下:

$$G = \begin{cases} -\log_2^{\frac{f(t)}{N}} & t \text{ 为简单术语} \\ -\sum_{w \in t} \log_2^{\frac{f(w)}{N}} & t \text{ 为复合术语} \end{cases} \quad (1)$$

式中,N 为语料库中总的文档数量,$f(.)$ 为一个词语的文档频次。如果一个术语是简单术语,也就是说它不包含其他术语或词语,那么其泛化度等于其信息量;如果一个术语是复合术语,即它由简单术

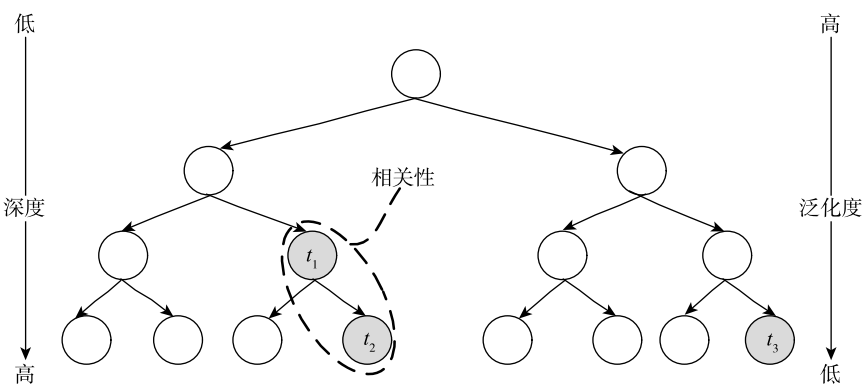

图1 术语的泛化度和相关性

语组成,或由简单术语与一般词语组成,那么术语的泛化度为其各个组合部分的信息量的累加。

设术语 t_1 的泛化度值为 G_1,术语 t_2 的泛化度值为 G_2,为了表示两个术语之间的泛化差异程度,定义两个术语的泛化度的差值的绝对值为术语对(t_1,t_2)的泛化度差值,泛化度差值越大,两个术语之间存在上下层次关系的可能性越大,计算公式如下:

$$D = |G_1 - G_2| \tag{2}$$

2.2 术语的相关度

在一个层级关系的概念体系中,两个具有直接上下位关系的概念具有较高的语义相似度,一般来说它们在语料库中存在较高的重现率[17]。例如,假设“信息资源”、“信息资源管理”和“政务信息资源管理”,相对而言,从概念层次上,“信息资源管理”较“信息资源”和“政务信息资源管理”接近。相比而言,在概念树上,层级位置接近的两个概念具有较高的关联性,深度相同但层级位置不接近的两个术语则关联性较低,如图 1 中的 t_1 和 t_2 具有较高的关联性,而 t_1 与 t_3 深度虽然相差 1,但语义关联性不强。

在一个概念层级关系中,验证两个概念之间是否接近,可以看看它们的相关程度,如果两个概念处于不同的子树上,则其不相关程度较高,反之,处于一个子树上、位置非常接近的两个概念,其相关程度较高。互信息是香农建立的信息论中一个重要的概念,可以直观地地衡量变量之间的依赖程度,常用来定义随机变量之间的相关和联系[19],因此,可以采用互信息来测量两个概念的关联性。在实际中,常用点互信息来代替互信息[20]。这里借鉴点互信息公式来定义术语的相关度如下:

$$association(t_i,t_j) = \left|\log \frac{tf_{ij}/N}{(tf_i/N)\times(tf_j/N)}\right| \tag{3}$$

式中,N 是语料库包含的论文数量,tf_i 表示术语 t_i 的文档频率,tf_j 表示术语 t_j 的文档频率,tf_{ij} 表示术语 t_i 和 t_j 共同出现的论文数量。

3 术语层次关系抽取方法

术语层次关系抽取是一个大规模的计算工作,单靠人工是难以完成的工作。对于数量规模为 n 的术语,术语的组合存在 $n(n-1)/2$ 种情况,当 n 非常大时,这个关系的增长是 2 次幂级变化。例如,假设有 1000 个术语,则存在的组合就有499 500种,所以通过机器来查找和筛选可能存在的关系是非常必要的。

3.1 词形规则特征模板

词形规则特征分析是一种模式匹配方法,它根据术语词对的构词特征进行术语之间关系的判别[21]。特征模板是一组启发式规则,包含类属关系模板、部分整体关系模板和其他模板。用 A、B、C 表示术语中包含的词语,按照两个术语在词形上的相同之处,采用表 1 中张巍定义一组特征模板[15],来判定两个术语之间的层次关系。

一般情况下,在表 1 中定义的术语关系模板中,满足 T1 模板的两个术语 C 和(A + C),可以确定术语 C 是术语(A + C)的上位概念(类),而术语(A + C)是术语 C 的下位概念(属);满足 T3 模板的两个术语 C 和(C + B),可以确定术语 C 是术语(C + B)的整体概念,而术语(C + B)是术语 C 的部分概念;同样,满足 T5 模板的两个术语也可以确定它们之间的 IS-A 关系。但对于 T2 模板和 T4 模板,如引言所述,则不能肯定两个术语之间一定存在层次关系,仍需判定其是否存在 IS-A 和 PART-OF 关系。

3.2 术语层次关系抽取原理

提出的术语层次关系抽取方法的原理见图 2。

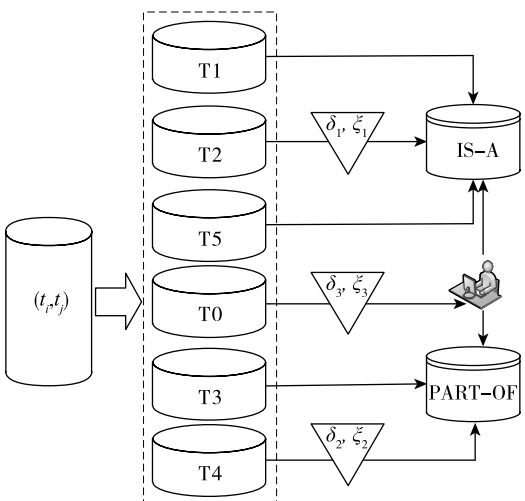

图 2 术语层次关系抽取原理示意图

首先将文本中抽取的术语两两配对,形成术语对列表,用(t_i,t_j)表示术语对。

对每一个术语对,采用最大公共子串算法(Longest Common Substring, LCS)[22]求它们的最大公共字符串。如果最大公共字符串同时发生在左侧或右侧(边缘),则进入模板匹配判断,采用表 1 定

义的模版规则进行判定。根据匹配模板的情况,将术语对分为6种,其中,T1～T5表示匹配对应模板的术语对,T0表示不匹配任何模板的术语对。

匹配T1、T5模板的术语对直接判定为IS-A关系。匹配T2模板的术语对经过泛化度差值阈值δ_1和相关度阈值ξ_1过滤器进行筛选,泛化度差值和相关度同时高于设定阈值的术语对被判定为IS-A关系,而不满足此条件的术语对被判定为不存在IS-A关系。

相似地,匹配T3模板的术语对直接判定为PART-OF关系。匹配T4模板的术语对经过泛化度差值阈值δ_2和相关度阈值ξ_2过滤器进行筛选,泛化度差值和相关度同时高于设定阈值的术语对被判定为PART-OF关系,而不满足此条件的术语对被判定为不存在PART-OF关系。

对于不匹配任何模板的T0术语对,则需要通过泛化度差值阈值δ_3和相关度阈值ξ_3过滤器进行筛选。泛化度差值和相关度同时高于设定阈值的术语对被认为存在层次关系,并提交给领域专家进行手工判断,由领域专家决定一个术语对是否存在层次关系,如果存在层次关系,则进一步判定是IS-A还是PART-OF关系。

3.3 阈值确定方法

提出的术语层次关系抽取涉及三组阈值(δ_1,ξ_1)、(δ_2,ξ_2)、(δ_3,ξ_3),δ表示泛化度差值阈值,ξ表示相关度阈值。

泛化度差值阈值δ_1和相关度阈值ξ_1由匹配T 1和T5模板的术语对确定。在实验中可以将T1和T5模板术语对的泛化度差值、相关度进行排序和分组,选择每一组的端点值判定抽取的效果,进而确定选择的阈值。

泛化度差值阈值δ_2和相关度阈值ξ_2由匹配T3模板的术语对确定。相似的,在实验中可以将T3模板术语对的泛化度差值、相关度进行排序和分组,选择每一组的端点值判定抽取的效果,进而确定选择的阈值。

泛化度差值阈值δ_3和相关度阈值ξ_3的确定需要考虑到模板中判定存在层次关系的术语对。实验中,可以将模板抽取的术语对的泛化度差值、相关度分别按照从大到小的顺序进行排序,然后按数据规模的一定比例来选取阈值点进行抽取效果的判定。例如,可以选取泛化度差值排序时前20%的术语对的最小值作为泛化度差值阈值,相关度排序时前20%的术语对的最小值作为相关度阈值。

3.4 抽取效果评价

采用正确率指标来测量抽取的效果,定义见公式:

$$\text{正确率} = \frac{\text{抽取正确的数量}}{\text{总抽取数量}} \times 100\% \qquad (4)$$

4 实验结果

从万方数据获取信息资源管理领域的科技论文数据,采用文献[23]提出的方法从文本中抽取术语,经过加工处理后,共抽取2930个术语。将抽取的中文术语两两组合形成中文术语对,计算了所有术语的泛化度指标,所有术语对的泛化度差值和相关度指标。

4.1 阈值选择实验

为了分析阈值选择对实验结果的影响,以便选择合适的阈值,采用随机方法选择了1300对术语作为实验数据,其中T2模板100对,T4模板200对,不符合任何模板的1000对,人工判断了这些术语对是否存在层次关系。将T1和T5模板确定的IS-A关系、以及T3模板确定的PART-OF关系的术语对泛化度差值和相关度值从最小值到最大值分为5个区间,包含最小值、最大值和均值共确定7个阈值,然后分别计算T2模板和T4模板在每一个阈值下的正确数和正确率,其中T2模板阈值由T1和T5模板确定,T4模板阈值由T3模板确定,结果见表2。可以看出泛化度差值阈值和相关度阈值越大,正确率越高,但是随着阈值的提高,抽取到的数量在减少,在阈值过高时抽取到的数量变为0,所以阈值的选择要在抽取的数量和正确率之间进行平衡,相对而言平均值是一个较好的选择。

为了确定不符合模板匹配规则的术语对的阈值,将符合T1、T3和T5模板术语对,以及T2和T4模板下判定为正确的术语对按照泛化度差值和相关度从大到小的顺序排列,分别取10%、20%、30%、40%和50%记录的泛化度差值最小值和相关度值最小值作为阈值,对随机选择的1000条不符合模板的术语对正确率进行了计算,结果见表3。随着泛化度阈值从大到小的变化,正确率在20%时达到最大值,但并没有发生单调变化的情况,而正确率随着相关度阈值的变小呈现不断下降的变化,10%相关度阈值对应的正确率为空是因为相关度阈值太高而没有满足条件的术语对。

表 2　T2 和 T4 模板在不同阈值下的正确率

		最小值	阈值 1	阈值 2	阈值 3	阈值 4	最大值	均值
T2	δ_1	0.2650	2.4522	4.6394	6.8266	9.0138	11.2012	2.2700
	正确数	33	4	1	0	0	0	5
	正确率	37.08	50.00	100.00				55.56
	ξ_1	0.3824	1.8903	3.3982	4.9061	6.4140	7.9220	2.6395
	正确数	35	20	5	0	0	0	11
	正确率	38.89	42.55	62.50				42.31
T4	δ_2	0.2650	2.2025	4.1400	6.0775	8.0150	9.9525	2.0923
	正确数	52	9	0	0	0	0	9
	正确率	29.71	27.27					25.00
	ξ_2	0.3824	1.8738	3.3652	4.8566	6.3480	7.8394	2.5690
	正确数	55	35	15	1	0	0	24
	正确率	31.25	33.65	46.88	100.00			40.00

表 3　不符合模板术语对在不同阈值下的正确率

	10%	20%	30%	40%	50%
δ_3	9.33	11.27	8.96	10.15	9.14
ξ_3		20.00	18.75	16.35	14.36

4.2　抽取结果

首先采用词形规则方法对所有可能的术语对进行了模板匹配。采用 T1 和 T5 模板抽取到的术语对共 476 个，可以确定为 IS-A 类关系，而采用 T2 模板抽取到的术语对共 942 个，术语关系可能是 IS-A 类，但仍需要进一步判断；采用 T3 模板抽取到的术语对共 515 个，可以确定为 PART-OF 类关系，而采用 T4 模板抽取到的术语对共 2497 个，可能是 PART-OF 关系，仍需进一步判定；不匹配任何模版的术语对（T0）共 143 431 个，其关系未定。表 4 列出了各匹配模板的术语对数量：

表 4　各模板抽取到的术语对数量

模板类型	匹配术语对数量
T1	471
T2	942
T3	515
T4	2 497
T5	5
T0	143 431

采用 T1 和 T5 模板抽取的 IS-A 关系类型的术语对的平均泛化度差值和平均相关度作为阈值，δ_1 为 2.2700，ξ_1 为 2.6395 ，采用该阈值对 T2 模板下的术语对关系进行判定，高于两个阈值的术语对判定为 IS-A 关系，小于阈值的则判定为不存在 IS-A 关系，共确定了 48 对术语。采用 T3 模板下术语对的平均泛化度差值和平均相关度值作为阈值，δ_2 为 2.5690，ξ_1 为 2.0923，对 T4 模板下的术语对关系进行判定，将相关度和泛化度差值均大于或等于设定阈值的术语对判定为 PART-OF 关系，不满足条件的则判定为不存在 PART-OF 关系，共确定了 145 对 PART-OF 关系的术语。为了对不满足模板的术语对进行层次关系判定，分别取前 20% 层次关系术语对泛化度差值指标和相关度指标从大到小排序的最小值作为阈值，δ_3 取 3.2716，ξ_3 取 4.1391，凡是相关度和泛化度差值同时大于等于阈值的术语对判定为存在层次关系，不满足阈值条件的则判定不存在层次关系，共获取 122 对术语。

经过以上处理后，共抽取具有层次关系的术语对 1306 个，其中 T1 模版 471 个，T2 模板 48 个，T3 模板 515 个，T4 模板 145 个，T5 模版 5 个，不匹配任何模板的 T0 术语对 122 个。将抽取到的 T2、T4 和 T0 术语对提交给领域专家进行判定。经过评估后，匹配 T2 模板的术语对有 38 个存在 IS-A 层次关系，匹配 T4 模板的术语对 127 个存在 PART-OF 关系，不匹配任何模板的术语对大多数有较强的语义关联关系，只有 3 个存在 IS-A 关系，49 个存在 PART-OF

关系。这样,共获取 1208 个具有层次关系的术语对,其中 IS-A 关系的术语对 517 个,PART-OF 关系的术语对 691 个,抽取结果及各模版下的正确率见表 5。

表 5　各模板下最终抽取的层次关系数量分布

模板	抽取数	IS-A	PART-OF	正确率(%)
T1	471	471	0	100.00
T2	48	38	0	79.17
T3	515	0	515	100.00
T4	145	0	127	87.59
T5	5	5	0	100.00
T0	122	3	49	42.98
总计	1306	517	691	92.50

在判定术语层次关系后,可以得到一些术语层次关系图。图 3 是以“信息资源管理”术语为根结点的 IS-A 层次关系图:

图 4 是以“信息资源管理”术语为根结点的 PART-OF 层次关系图:

对抽取不正确的术语关系进行了分析。虽然设置的泛化度差值和相关度阈值条件过滤掉了一部分不存在层次关系的术语对,但仍无法避免错误的抽取,这可能是由于采用互信息来测量相关度引起的。例如抽取的(管理技术,高新技术)术语对并不具有层次关系,但两者同时具有较高的泛化度差值和相关度;两者的泛化度差值是 2.4035,这是由于在信息资源管理领域中“管理”远比“高新”常用,所以“管理技术”的泛化度远高于“高新技术”,这是正确的;“管理技术”出现的文档频次是 43,高新技术出现的文档频次是 25 次,两词共献的频次是 2,而两者的相关度高达 2.9394,产生这个问题的原因在于互信息本身的特点,因为互信息在其他条件相等的情况下,由低频率词组成的二元组的互信息要大于高频率词组成的二元组[20]。

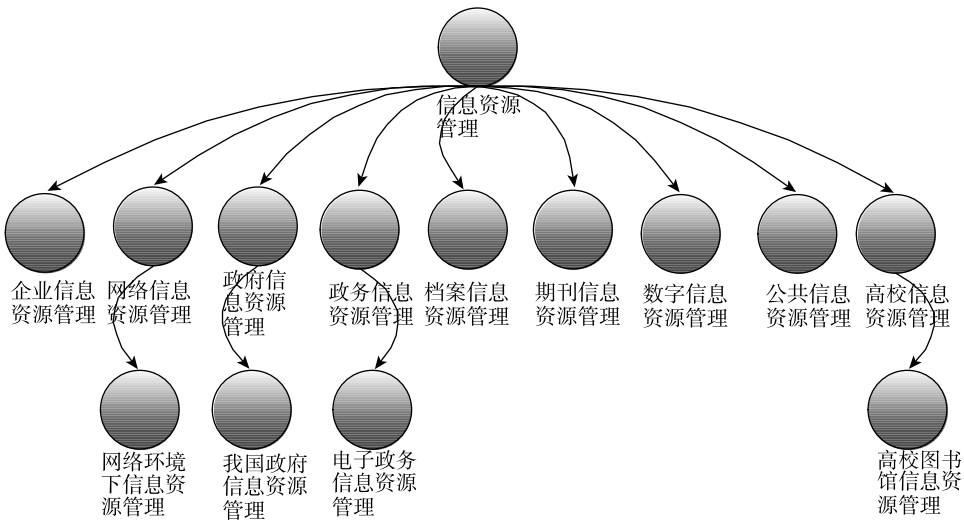

图 3　信息资源管理 IS-A 术语关系层次图

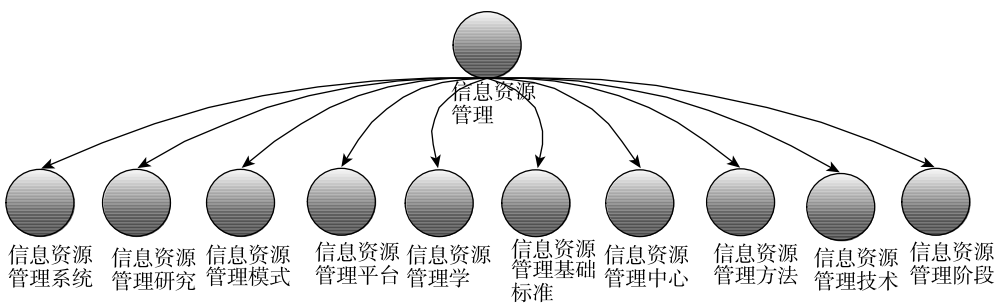

图 4　信息资源管理 PART-OF 术语关系层次图

5 总 结

研究了从科技文献文本中抽取 IS-A 和 PART-OF 层次关系的术语关系抽取方法,提出了基于词形规则模板的方法。词形规则模板利用了多词术语的 head 和 modifier 特征,采用 5 类模板来确定边缘词汇相同的术语之间的层次关系;提出了泛化度指标来测度术语的泛化,用于确定两个术语在概念层次树的上下位关系;提出相关度指标来测度两个术语之间的相关度,用于确定两者在语义概念上是否接近。

采用提出的方法对信息资源管理领域论文文本中的术语层次关系抽取进行了实验。共抽取中文术语 1306 对,经过人工评估后,最终选定 1208 对,其中具有 IS-A 层次关系的 517 对,具有 PART-OF 层次关系的 691 对。实验结果证明本研究提出的术语层次关系抽取方法是有效的,可以从语料库中抽取到一定数量的具有 IS-A 或 PART-OF 层次关系的领域概念术语对。从抽取效果上来看,总体上达到了较高的正确率。但是还应该看到,提出方法实现较好的正确率主要得益于模板抽取的结果,在不符合模板匹配的术语对层次关系抽取上效果仍然较差,抽取正确率只有 42.98%,另外对不正确的抽取进行了初步分析,发现采用互信息测量相关度可能带来一些抽取错误,这个问题需要在未来进行解决。

参 考 文 献

[1] 何琳. 基于多策略的领域本体术语抽取研究 [J]. 情报学报, 2012, 31(8): 798-804.

[2] 孙霞, 王小凤, 董乐红,等. 术语关系自动抽取方法研究 [J]. 计算机科学, 2010, 37(2): 189-191,215.

[3] Boguraev B,Kennedy C. Applications of term identification technology:domain description and content characterization [J]. Natural Language Engineering,1999,5(1):17-44.

[4] ISO. Terminology work-principles and methods[S]. 2009.

[5] 贾秀玲, 文敦伟. 一种本体学习中分类关系提取方法的研究 [J]. 计算机技术与发展, 2007, 17(10): 31-33,36.

[6] Mark Sanderson,Croft Bruce. Deriving concept hierarchies from text [C]// Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999: 206-213.

[7] Brian Roark, Charniak Eugene. Noun-phrase Co-occurrence Statistics for Semi-automatic Semantic Lexicon Construction [C]//36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1998: 1110-1116.

[8] Hui Yang, Callan Jamie. A Metric-based Framework for Automatic Taxonomy Induction [C]//Joint conference of the 47th annual meeting of the Association for Computational Linguistics and 4th international joint conference on natural language processing of the Asian Federation of Natural Languages processing 2009 (ACL-IJCNLP 2009),2009:271-279.

[9] Hearst M A. Automatic acquisition of hyponyms from large text corpora [C]. 1992: 539-545.

[10] Ellen Riloff,Shepherd Jessica. A Corpus-based Approach for Building Semantic Lexicons [C]. 1997: 117-124.

[11] Marti A. Hearst. Automated discovery of WordNet relations [A]//Christiane Fellbaum ed. To Appear in WordNet: An Electronic Lexical Database and Some of its Applicationsed[M]. MIT Press, 1998:131-151.

[12] 刘威. 基于中文文本的本体构建方法研究 [D]. 哈尔滨: 哈尔滨工程大学, 2008.

[13] Gregory Grefenstette. SQLET : Short query linguistic expansion techniques: Palliating one or two-word queries by providing intermediate structure to text [C]. 1997: 500-509.

[14] William A Woods. Conceptual Indexing: A Better Way to Organize Knowledge[R]. CA, UAS Sun Microsystems, Inc. Mountain View, 1997.

[15] 张巍, 于洋, 游宏梁. 面向词汇知识库自动构建的概念术语关系识别 [J]. 现代图书情报技术, 2009 (11): 10-16.

[16] Pum-Mo Ryu, Choi Key-Sun. An information-theoretic approach to taxonomy extraction for ontology learning [A]. In:Buitelaar P ed. ,Ontology Learning from Text: Methods,Evaluation and Applicationsed[M]. Lansdate PA. USA: IOS Press, 2005.

[17] 陈珂. 构造领域本体概念关系的自动抽取 [D]. 上海: 上海交通大学, 2008.

[18] 何琳. 领域本体的关系抽取研究 [J]. 现代图书情报技术, 2008(4): 35-38.

[19] Church K, Hanks K. Word association norms, mutual information and lexicography[J]. Computational Linguistics, 1990, 16(1): 22-29.

[20] 张勇. 中文术语自动抽取相关方法研究 [D]. 华中师范大学, 2006.

[21] Paola Velardi, Fabriani Paolo, Missikoff Michele. Using text processing techniques to automatically enrich a domain ontology [C]//Proceedings of the international conference on Formal Ontology in Information Systems (FOIS'01), 2001: 270-284.

[22] 潘虹, 徐朝军. LCS 算法在术语抽取中的应用研究 [J]. 情报学报, 2010, 29(5): 853-857.

[23] 韩红旗, 朱东华, 汪雪锋. 专利技术术语的抽取方法 [J]. 情报学报, 2011, 30(12): 1280-1285.

(责任编辑 马 兰)

基于词形规则模板的术语层次关系抽取方法

作者： 韩红旗， 徐硕， 桂婕， 乔晓东， 朱礼军， 安小米， Han Hongqi， Xu Shuo， Gui Jie ， Qiao Xiaodong， Zhu Lijun， An Xiaomi

作者单位： 韩红旗,徐硕,桂婕,乔晓东,朱礼军,Han Hongqi,Xu Shuo,Gui Jie,Qiao Xiaodong,Zhu Lijun(中国科学技术信息研究所,北京,100038)， 安小米,An Xiaomi(数据工程与知识工程教育部重点实验室(中国人民大学)中国人民大学信息资源管理学院,北京,100872)

刊名： 情报学报 ISTIC PKU CSSCI

英文刊名： JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC ANDTECHNICAL INFORMATION

年，卷(期)： 2013,32(7)