面向维基百科的领域知识演化关系抽取

高俊平[1)] 张 晖[2)] 赵旭剑[1)] 杨春明[1)] 李 波[1),3)]

[1)](西南科技大学计算机科学与技术学院 四川 绵阳 621010)

[2)](西南科技大学教育信息化推进办公室 四川 绵阳 621010)

[3)](中国科学技术大学计算机科学与技术学院 合肥 230027)

摘 要 互联网下同一领域中不同知识概念间存在多种关系,其中演化关系对于用户学习和理解领域知识,梳理领域知识的前序和后续逻辑关系具有重要意义,然而网络数据的多样和无序使用户难以准确有序地获取领域知识关系.针对该问题,提出一种面向中文维基百科领域知识的演化关系抽取方法,利用语法分析特征,挖掘演化关系模式,构建演化关系推理模型,采用基于句子层面的关系抽取算法识别领域知识演化关系,最后在真实的维基百科数据集上对该文方法进行了性能评测.实验表明,该方法具有较高的关系抽取准确率和召回率,能有效地抽取出维基百科中领域知识的演化关系.同时,基于实验抽取结果构建知识图谱,能有效挖掘领域学科下知识集合的演化体系,识别重难点知识,对学科建设以及相关课程教学具有一定的指导意义.

关键词 领域知识;维基百科;演化关系;关系抽取;条件随机场;社会媒体

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2016.02088

Evolutionary Relation Extraction for Domain Knowledge in Wikipedia

GAO Jun-Ping[1)] ZHANG Hui[2)] ZHAO Xu-Jian[1)] YANG Chun-Ming[1)] LI Bo[1),3)]

[1)](*School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan* 621010)

[2)](*Educational Informationization Office, Southwest University of Science and Technology, Mianyang, Sichuan* 621010)

[3)](*School of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027)

Abstract There are many relations among different concepts in the same field of knowledge from Internet. The evolutionary relation is very important for users to learn and understand the domain knowledge as well as sorting out the successive logic relationship between two different concepts. However, the diversity and disorder of the network data made it difficult for users to obtain relations among domain knowledge accurately and orderly. Aiming at this issue, we propose an evolutionary relation extraction method for domain knowledge in Chinese Wikipedia. Firstly, we construct a reasoning model for evolutionary relation using different syntax features as well as patterns discovery for evolutionary relation, and then detect the evolutionary relation for domain knowledge taking advantage of a sentence level relation extraction algorithm. Finally, we evaluated the method on the real Wikipedia data set in this paper. Experiments show that our method has higher precision and recall than the existing models and our proposal is effective against evolutionary relation extraction for domain knowledge in Wikipedia. Moreover, based on the experimental results, we construct a knowledge map in one field, which can represent the

收稿日期:2015-10-20;在线出版日期:2016-03-05.本课题得到四川省教育厅资助项目(14ZB0113)、西南科技大学博士基金(12zx7116)和赛尔网络下一代互联网技术创新项目(NGII20150510)资助.**高俊平**,男,1991年生,硕士研究生,主要研究方向为信息抽取.E-mail:gjp0228@qq.com.**张 晖**(通信作者),男,1972年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为文本挖掘、知识工程.E-mail:zhanghui@swust.edu.cn.**赵旭剑**,男,1984年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为中文信息处理、Web信息检索.**杨春明**,男,1980年生,硕士,副教授,中国计算机学会(CCF)会员,主要研究方向为文本挖掘、知识工程.**李 波**,男,1977年生,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究方向为信息过滤、信息安全.

evolution structure of the knowledge-object collection and identify important and difficult points of knowledge effectively. Therefore, this method has a certain guiding significance to the subject construction and the related course teaching.

Keywords domain knowledge; Wikipedia; evolutionary relation; relation extraction; conditional random fields; social media

1 引　言

随着信息技术与网络的发展，互联网已成为一个包含大量非结构化、异构数据的信息源，以维基百科为代表的网络知识服务平台日益崛起，互联网知识数据急剧增长. 然而，网络知识数据在快速积累的过程中具有数量庞大、内容丰富、类型多样、动态性强、无序性大的特点[1]，给用户准确、有序地获取领域知识带来了巨大的挑战. 因此，对互联网数据进行规则整理，挖掘知识概念实体关系，对于面向互联网的领域知识研究具有重要研究意义[2-4]. 然而目前的研究大多关注领域知识的挖掘与抽取，忽视了领域知识相互关联的研究[5-8]. 同一领域中不同知识概念存在多种关系，其中概念间的演化关系反映了该领域知识的演变发展过程，对于用户学习和理解领域知识，梳理领域知识的前序和后续逻辑关系具有重要意义. 例如，从句子“支持向量机的理论基础来自于统计学习理论”可以抽取出“支持向量机”和“统计学习”之间的演化关系，阐明两者存在的理论关联，为人们学习和理解“支持向量机”相关理论提供准确、全面的领域前序知识. 针对这一实际需求，本文提出了一种面向维基百科的领域知识演化关系抽取方法，通过句法分析构建领域知识演化关系推理模型，将关系抽取转化为序列标注问题，利用条件随机场模型对中文维基百科进行领域知识演化关系抽取. 与已有工作相比，本文的主要贡献在于：

(1) 建立了领域知识演化关系推理模型(见第 3 节)，为不同结构、不同语义的领域知识关系表达建立准确的句法分析机制，利用知识概念之间的语义角色关系设计领域知识演化关系模式.

(2) 提出了基于条件随机场(Conditional Random Fields，CRF)模型的领域知识演化关系抽取方法(见第 4 节)，对于不同的演化关系模式，建立了统一的关系抽取理论模型. 在中文维基百科数据集上的实验表明，该算法较同类抽取模型具有更好的实验性能(见第 5 节)，能有效地发现领域知识之间的演化关系.

2 相关工作

随着网络资源的日益丰富，面向互联网的领域知识研究受到不少学者的广泛关注. 人们主要通过定义不同的领域知识表示方式，以及设计自动化的知识获取方法发现海量互联网数据中的领域知识[9-10]. 此外，洪俊[11]提出基于 Deep Learning 的领域概念抽取算法，主要分为训练和测试两个模块，首先通过训练数据学习得到深度网络模型，然后在测试模块中利用上一步训练得到的深度模型对测试数据进行自动分类识别，对于分类结果，通过人工审核的方式最终获取正确的领域概念集. 董丽丽等人[12]则建立一种结合语义相似度和改进近邻传播算法的领域概念自动获取方法，通过互信息进行合成词提取，使用对数似然比的策略避免遗失低频词，利用 HowNet 和余弦相似度识别术语间同义词，采用改进的近邻传播算法获取领域概念集合. 然而，领域知识并不是固定不变的，受某些情况的影响会产生新的概念(知识)，这就是概念的演化，也可以认为它是概念漂移[13](Concept Drift)的一种表现形式. 传统的有关概念漂移的研究工作[14-17]主要探讨对于概念发生漂移过程的监测方法，关注的焦点是随着时间的变化，概念发生的变化；同时，为了准确有效地挖掘数据流中的概念，文献[18]提出一种基于时间衰减模型和闭合算子的数据流闭合模式挖掘方式 TDMCS 算法，采用时间衰减模型来区分滑动窗口内的历史和新近事务权重，使用闭合算子提高闭合模式挖掘的效率，设计使用最小支持度-最大误差率-衰减因子的三层架构避免概念漂移. 然而就概念演化而言，关注的则主要是概念间的前继与后继关系，即概念的发展基础以及概念的后续发展.

关系抽取(Relation Extraction，RE)作为信息抽取(Information Extraction，IE)的关键技术之一，近年来成为 Web 数据挖掘的研究热点[19-23]. 以维基百科为代表的互联网知识服务平台，由于其拥有数据丰富、结构规范、领域性强等特点，在关系抽取相关领域发挥着重要作用. 张苇如等人[24]提出了从人

工标注知识体系到维基百科实体映射的方法获取关系实例,通过利用维基百科的结构化特性,解决了实体识别的准确性问题,产生了准确和显著的句子实例. Yu 和 Lam[25]提出了一种新型的基于马尔可夫逻辑网络的多特征的维基百科关系抽取方法,是一种针对维基百科数据的概念上下位关系抽取模型. 汤青等人[26]利用基于"是一个"模式匹配的原则进行上下位关系抽取研究,把上下位关系的抽取问题转化为上位概念和下位概念的抽取. Caraballo[27]通过连接词和同位语特征,利用上下文中的名词特征的连接关系或同位关系来构造出名词特征向量,使用聚类的方法获取名词间的上下位关系. 文献[28]提出一种基于条件随机场的领域术语上下位关系获取方法,结合百科名片中结构化、制式化的语言表达形式,通过统计分析,构建适用于通用模型的特征词词典,再在词和词性特征的基础上结合特征词词典内容和标点符号信息,利用 CRF 模型抽取上下位关系. 同时,研究学者[29]认为可以把领域概念间的关系分为上下位关系、等同关系、与关系、交叉关系、或关系、非关系、矛盾关系、因果关系等关系类型,同时对不同的关系都作了相关的定义与研究,例如将内涵完全相同的两个或两个以上概念间的关系称为等同关系,然后根据 OWL(Web Ontology Language)[30]语言使用的 owl:equivalentClass 标签表示概念间的等同关系,却对于知识之间存在的概念演化关系缺乏考虑.

总的来说,已有方法对于领域知识关系抽取的研究主要基于文本结构以及简单的语法分析,这种机制对于具有复杂结构和语义表达的中文领域知识演化关系抽取存在明显不足,两个概念间的句法多样性大大影响抽取算法性能. 因此,本文提出一种基于演化关系推理模型的领域知识演化关系抽取方法,通过深层句法分析,构建领域知识演化关系模式,利用 CRF 模型学习关系模式特征,抽取领域知识演化关系. 实验结果表明,考虑深层句法特征的抽取方法较传统方法具有更高的准确性,更适合中文领域知识演化关系抽取.

3 演化关系推理模型

领域知识间的演化关系包含两个对象:前置概念实体(pre-concept)和后置概念实体(post-concept),分别表示具有演化关系的前序知识和后续知识. 为此,本文针对演化关系作出如下定义.

定义 1. 如果对于给定概念实体 A,B 以及满足演化关系约束条件的句子 $S(A\langle$约束条件$\rangle B)$,则演化关系表示为 $evolution(A,B)$,其中,A 称为 B 的前置概念,B 称为 A 的后置概念.

对于同时包含两个领域知识概念实体的句子,根据句子中的谓词来判断其是否满足演化关系中约束条件. 例如,句子"支持向量机的理论基础来自于统计学习理论",谓词"来自于"是具有演化关系的特征词;再如,句子"随后,由三定律得到了机械伦理学",谓词"得到"同样是具有演化关系的特征词;因此这两个句子均满足演化关系约束条件. 特征词的不同反映了演化关系顺序上的不同(如表 1 所示).

表 1 演化关系特征词

类别	特征词	实例	关系表达式
正序关系	发展、衍生、得到等	x 衍生出 y	$evolution(x,y)$
逆序关系	来自、基础、继承等	x 来自于 y	$evolution(y,x)$

3.1 演化关系模式

具有演化关系的特征谓词在语句中起到了前置概念与后置概念的语义关联作用,因此,利用这一性质构建演化关系模式,如下所示:

$\{$演化关系模式:$\langle \#concept_1\rangle\langle$ #停用词$\rangle$
$\langle$特征谓词$\rangle\langle \#concept_2\rangle\}$

模式中"$concept_1$"和"$concept_2$"指的是句子中出现的领域知识概念实体,"$\langle \#concept_1\rangle$"、"$\langle \#concept_2\rangle$"指的是包含概念实体的任意字符串,停用词指的是与内容无关的词,如"的、地"等. 对于给定句子 S,如果它的句子结构与上述模式结构相匹配,则认为 S 中包含的两个概念实体具有演化关系. 因此,针对具有演化关系的句子,将概念间演化关系抽取问题转化为演化关系中的前置概念与其相对应的后置概念的抽取问题. 显然,特征谓词在句子 S 中起到谓语成份,前置概念和后置概念分别作为特征谓词的宾语(或主语)和主语(或宾语),S 中包含前置概念的部分内容标记为 $preS$,相应包含后置概念的部分标记为 $postS$.

然而并非所有具有演化关系的句子都满足上述模式,如"有些图像压缩演算法就是以奇异值分解为基础",显然,概念"图像压缩演算法"与"奇异值分解"具有知识演化关系,但该句子的句法结构却不符合演化关系模式. 因此,将中文语句分为两类:句子结构符合演化关系模式的称为简单句(简单模式);不符合关系模式但满足演化关系约束条件的句子称为复杂句(复杂模式). 对于两类结构差异的中文语

句，分别采用不同的推理方法进行概念实体间的演化关系分析，具体内容将在下面讨论.

在本文的研究过程中，首先希望通过机器学习的方法自动进行简单模式与复杂模式的分类，然后针对不同模式进行不同的演化关系推理，最后根据演化关系推理方法进行演化关系抽取. 但是演化关系模式分类的实验结果并不理想，这对演化关系抽取会产生一定的影响. 为了验证真实的评测效果，本文针对模式分类结果，采用模板匹配的方法进行演化关系抽取实验，实验结果表明，仅仅依据模板匹配的方式进行演化关系抽取，无法获得良好的效果(见5.2节). 因此本文对不同的演化关系模式进行不同的演化关系推理，然后对这两种演化关系模式统一进行基于CRF监督式学习的演化关系抽取.

3.2　演化关系推理

根据上述的定义，简单句的句子结构相对简单，而且一般只含有一个主谓结构，其中特征谓词作谓语，前置概念和后置概念分别充当宾语(或主语)成分和主语(或宾语)成分. 因此对于简单句中概念演化关系抽取主要是抽取演化关系中的前置概念以及与之对应的后置概念，后置概念从 *postS* 中获得，前置概念从 *preS* 中获得. 例如，对于简单句“支持向量机的理论基础来自于统计学习理论”，分析结果(如图1所示)可以发现具有演化关系的特征词“来自”作谓语，在 *preS* 中含有前置概念“统计学习理论”作演化关系句子 *S* 的宾语，在 *postS* 中含有后置概念“支持向量机”. 因而“支持向量机”与“统计学习”具有逆序演化关系，推理结果表示为 *evolution*(统计学习，支持向量机). 一般来说，对于简单句的演化关系抽取，主要是基于模式匹配的方法，先发现 *preS* 与 *postS*，再从两者之中获取前置概念与后置概念.

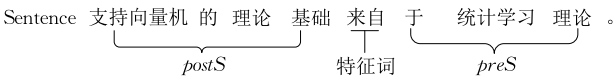

图1　简单句实例分析结果

但是通过实验数据的分析发现，对复杂句进行模式匹配不能准确地抽取句子中的演化关系，需要对句子结构进行深入分析. 因此，利用依存句法对复杂句进行句法分析，发现复杂句中的概念实体演化关系主要体现在五类句法结构：主谓关系(SBV)、动宾关系(VOB)、定中关系(ATT)、状中结构(ADV)以及左(右)附加关系(L|RAD)关系. 例如，对于复杂句“概率论是今天数理统计的基础”，从依存句法分析结果(如图2所示)可以发现具有演化关系的特征谓词“基础”在句法层面上作宾语成分(VOB)，而概念实体“概率论”与“数理统计”则都在依存句法分析中的主谓关系(SBV). 因此，根据概念实体与特征谓词的语法关系特征，可以推理概念实体之间的演化关系，注意到特征谓词“基础”充当“数理统计”的主谓关系(SBV)宾语，因而“概率论”与“数理统计”具有正序演化关系，推理结果表示为 *evolution*(概率论，数理统计).

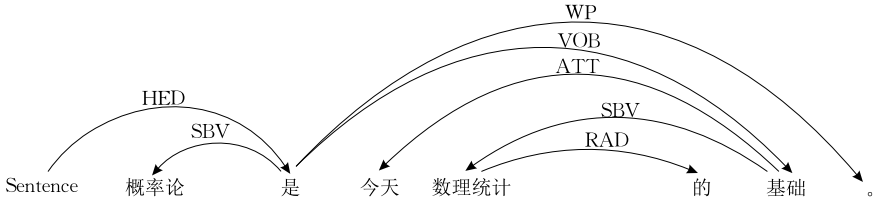

图2　复杂句实例依存句法分析结果

4　演化关系抽取方法

在本文中，使用条件随机场(Conditional Random Fields，CRF)模型抽取领域知识演化关系，它最早是在2001年由Lafferty等人[31]提出，是在最大熵模型和隐马尔可夫随机场基础上产生，是一种具有区分性的无向图模型，解决了隐马尔可夫链(Hidden Markov Model，HMM)在重叠和非独立特征上的困难[32].

CRF模型假设 $m_{1:N}$ 是一个显性序列(可观察的序列，例如文档中的词)，$k_{1:N}$ 是一个隐性状态(例如标注)，如图3所示. 线性CRF根据 m 和 k 构建了一个条件概率模型 $P(k|m)$. 模型定义如式(1).

$$P(k|m)=\frac{1}{Z(m)}\exp\Big\{\sum_i\sum_n\lambda_n f_n(k_{i-1},k_i,m,i)+\sum_i\sum_n\mu_n g_n(k_i,m,i)\Big\} \tag{1}$$

其中，$Z(m)$ 是归一化因子也称之为区分函数，使得 $P(k|m)$ 在标注序列上的和为1；$f_n(k_{i-1},k_i,m,i)$ 是状态转移函数；$g_n(k_i,m,i)$ 是状态特征函数.

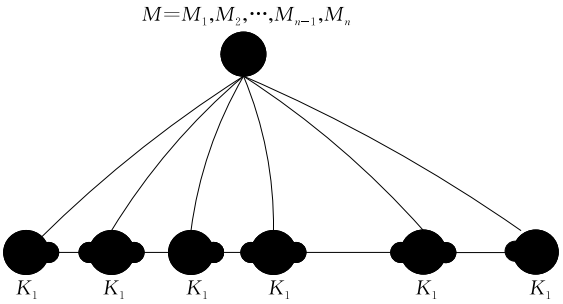

图 3 CRF 标注序列

4.1 CRF 句子层面关系抽取算法

概念实体间的关系可以用关系三元组$\{\varepsilon_1,R,\varepsilon_2\}$来描述,其中$\varepsilon_1$和$\varepsilon_2$分别指所识别的前置概念与后置概念,R 指概念实体间的关系类型(本文中特指演化关系 *evolution*).在这里本文提出基于 CRF 的句子层面上的关系抽取算法,利用该算法抽取出句子中包含的演化关系三元组对象.算法的基本流程如算法 1 所示.

算法 1. 关系抽取.

输入:S: a sentence

输出:*r_record*: a relation record $\{\varepsilon_1,R,\varepsilon_2\}$

1. BEGIN
2. $ArgumentList=SyntacticParse(S)$;
3. $FeatureList=FeatureSelection(ArgumentList)$;
4. $LabeledEntities=\mathrm{CRF}(ArgumentList,FeatureList)$;
5. RETURN $EntityRecognition(LabeledEntities)$;
6. END

抽取算法主要分为 4 步:(1)参数构建.利用 *SyntacticParse* 函数对句子成份进行处理,解析句法结构;(2)特征选取.利用 *FeatureSelection* 函数对每个输入的句子进行特征抽取;(3)利用 CRF 模型对句子成份进行序列标注,训练抽取模型;(4)获取演化关系三元组.

4.2 特征选择

在训练 CRF 模型时,针对特定的任务为模型的构建选择合适的特征集合以及特征函数的定义是重要的一项任务.在自然语言处理中,常用的特征主要有 POS 类型、命名实体识别、依存句法和语义角色,通过第 3 节中演化关系推理模型的分析,本文认为词的位置对抽取模型类型标注同样有一定的影响,因此将这五类特征作为特征集合(如表 2 所示)标注中文句子序列.不同的特征组合对演化关系的抽取会有不同的效果,具体实验结果见第 5 节.

特征集中的 POS 类型、命名实体识别、依存句法以及语义角色标注会有不同的可选值①,本实验所涉及到的主要特征值及其具体说明见表 3~表 6.

表 2 特征集

特征	说明	可选值
词位置	词所在句子中的位置	
POS 类型	词的词性	v,p,n,u 等(863 词性标注集)
命名实体识别	词的命名实体类型	Nh,Ni,Ns,O
依存句法	词与词在句子中的依存关系	SBV,VOB,ATT,POB 等
语义角色标注	语义角色标注	A0,A1,NULL 等

表 3 POS 类型可选值说明

标记	说明	示例
v	verb 动词	支持、统计
p	Preposition 介词	于、在
n	general noun 一般名词	理论、概率
u	auxiliary 助动词	的
wp	punctuation 标点	,。?
nt	temporal noun 时间名词	今天、明天
j	abbreviation 缩写	数理

表 4 命名实体识别可选值说明

标记	说明	示例
Nh	人名	邓丽君
Ni	机构名	外交部
Ns	地名	北京市
O	非命名实体	来自

表 5 依存句法可选值说明

标记	说明	示例
HED	核心关系	整个句子的核心
ADV	状中结构	非常美丽(非常←美丽)
ATT	定中关系	红苹果(红←苹果)
RAD	右附加关系	孩子们(孩子→们)
SBV	主谓关系	我送她一束花(我←送)
VOB	动宾关系	我送她一束花(送→花)
WP	标点符号	
COO	并列关系	山和海(山→海)
POB	介宾关系	在贸易区内(在→内)
CMP	动补结构	来自于(来自→于)

表 6 语义角色标注可选值说明

标记	说明	示例
A0	施事	支持向量机
A1	受事	统计学习
NULL	无语义角色	的

4.3 标记策略

利用 CRF 模型对句子成分进行序列标注主要采用了 LE(Left-Entity)、RE(Right-Entity)、RT(Relation-Type)、OT(Other)标签,并结合自然语言处理中常用的"BIE"标记模式,考虑到包含演化关系的句子主要分为简单句与复杂句,因此对不同的模式,有不同的标记策略.

① http://www.ltp-cloud.com

如 3.2 节中提到的简单句的演化关系抽取主要是抽取演化关系中的前置概念以及与之对应的后置概念，后置概念从 *post*S 中获得，前置概念从 *pre*S 中获得. 因此对于简单句的标记主要是标记 *post*S、*pre*S 与演化关系特征词. 图 4 给出了一个简单句的标记实例.

0	支持	v	O	−1	HED	A0	**RE-B**
1	向	p	O	6	ADV	A0	**RE-I**
2	量机	n	O	5	ATT	A0	**RE-E**
3	的	u	O	2	RAD	NULL	**OT**
4	理论	n	O	5	ATT	NULL	**OT**
5	基础	n	O	1	POB	NULL	**OT**
6	来自	v	O	0	VOB	NULL	**RT**
7	于	p	O	6	CMP	NULL	**OT**
8	统计	v	O	7	POB	A1	**LE-B**
9	学习	v	O	8	COO	A1	**LE-E**
10	理论	n	O	9	VOB	NULL	**OT**
11	。	wp	O	0	WP	NULL	**OT**

图 4　简单句标记

复杂句虽然不符合关系模式但满足演化关系约束条件，因此对它的标记主要结合句子结构，利用依存句法分析结果对句子进行标记. 如 3.2 节中提到的复杂句中的概念实体演化关系主要体现在 5 类句法结构：主谓关系(SBV)、动宾关系(VOB)、定中关系(ATT)、状中结构(ADV)以及左(右)附加关系(L|RAD)关系. 具有演化关系的特征谓词在句法层面上作宾语成分(VOB)，而概念实体则在依存句法分析中的主谓关系(SBV). 因此对于复杂句的标记主要是标记句子中的宾语成分(VOB)的特征谓词与主谓关系(SBV)的概念实体. 图 5 给出了一个复杂句的标记实例.

0	概率	n	O	1	ATT	A0	**LE-B**
1	论	v	O	2	SBV	A0	**LE-E**
2	是	v	O	−1	HED	NULL	**OT**
3	今天	nt	O	7	ATT	NULL	**OT**
4	数理	j	O	5	SBV	A1	**RE-B**
5	统计	v	O	7	ATT	A1	**RE-E**
6	的	u	O	5	RAD	NULL	**OT**
7	基础	n	O	2	VOB	NULL	**RT**
8	。	wp	O	2	WP	NULL	**OT**

图 5　复杂句标记

5　实　验

5.1　实验设置及评估标准

实验数据来自中文维基百科[①]，首先我们以“机器学习”词条为种子，通过爬虫爬取到十万个网页，然后通过人工过滤从十万个网页中筛选出与“机器学习”领域知识话题相关的 1000 个网页进行实验，对实验数据采用 10 倍交叉验证，9 份作为训练集，1 份作为测试集. 实验中采用 CRF++[②]工具包进行训练和测试.

评估标准采用的是信息检索领域中标准的评价指标：准确率(Precision)、召回率(Recall)和 F 值(F-score)进行评估. 准确率是指标注为类别 X 的实例中标注正确的比例，召回率是指类别为 X 的实例中标注正确的比例，而 F 值则是对准确率和召回率的整体衡量，可通过式(2)计算得到.

$$F=2\times P\times R/(P+R) \tag{2}$$

5.2　演化关系模式分类评价

根据 3.1 节的说明，领域知识的演化关系主要有简单模式与复杂模式两种，本文首先尝试通过机器分类的方法将这两种模式进行有效地区分，针对不同的模式结合不同的推理方法进行演化关系推理，基于机器学习的模式分类处理方法如图 6 所示.

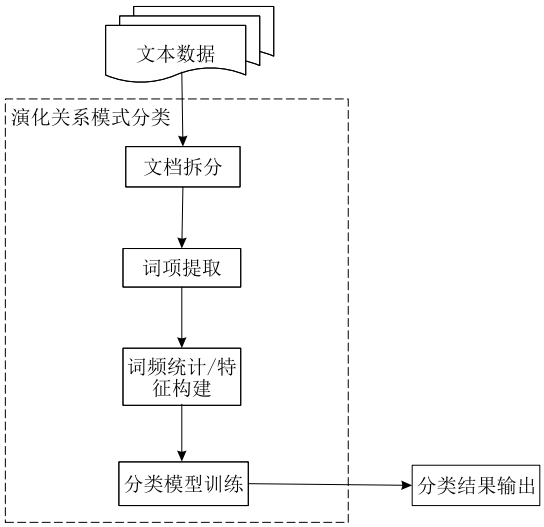

图 6　基于机器学习的模式分类

其中，本文采用的分类模型主要有两种：基于概率的朴素贝叶斯(Naïve Bayesian，NB)分类模型与基于特征的支持向量机(Support Vector Machines，SVM)分类模型. 基于机器学习的演化关系模式分类主要有 4 个步骤：

(1) 文档拆分. 对读入的文档进行预处理，根据常用的中文或英文符号(如。.??!!等)作为分句标识，将文档拆分为句子，以句子为基本单位进行演化关系模式分类.

(2) 词项提取. 针对(1)中处理后的数据进行中

① http://zh.wikipedia.org/
② http://sourceforge.net/projects/crfpp/

文分词处理，为构建 SVM 模型提取候选句子中的特征词，同时为构建 NB 模型提取句子中的每个词项.

(3) 词频统计/特征构建. 由于分类模型的不同，第三步工作分为两种情况. 针对朴素贝叶斯模型，主要统计各个词项的频次信息，而对于支持向量机模型，则主要完成特征(函数)的构建，这里我们对句子中的特征谓词通过计算其 TF-IDF(Term Frequency-Inverse Document Frequency)值来区分权重大小.

(4) 分类模型训练. 根据(3)中计算的最大后验概率以及多元特征函数输出分别训练 NB 分类模型和 SVM 分类模型，最后将分类结果输出.

本文的演化关系模式分类实验数据来自于 5.1 节所给出的数据(6 份作为训练集，4 份作为测试集)，评估标准同样采用了准确率(Precision)、召回率(Recall)和 F 值(F-score)，实验结果如图 7 所示.

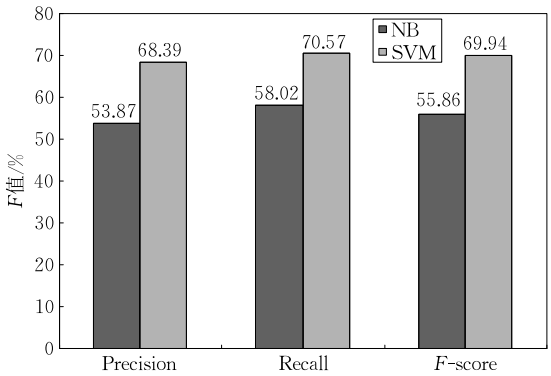

图 7 演化关系模式分类实验结果

实验结果表明基于支持向量机的演化关系模式分类模型优于基于朴素贝叶斯的演化关系模式分类模型. 这主要是由于朴素贝叶斯模型仅考虑假设数据样本的特征项之间相互独立，而支持向量机模型则综合考虑了数据样本中特征项之间的关联.

然而，无论是基于概率的朴素贝叶斯分类模型还是基于特征的支持向量机分类模型，两种分类模型在演化关系模式分类上的实验效果并不理想. 其中，朴素贝叶斯分类模型的 F 值仅为 55.86%，而支持向量机分类模型的 F 值略有提升，仍不足 70%，这对于不同演化关系模式(简单模式和复杂模式)的演化关系抽取噪声较大.

为进一步验证演化关系模式分类对演化关系抽取的影响，本文以支持向量机分类结果(如表 7 所示)作为关系抽取的实验数据(此处仅考虑简单模式)，同时以 3.1 节中描述的演化关系模式为基础构建演化关系抽取模板，利用模板匹配的方式进行演化关系抽取.

表 7 演化关系模式分类结果

模式类别	数量	实例
简单模式	104	计算机视觉问题理论基础是统计学最优化理论
复杂模式	213	变分法提供了有限元方法的数学基础
噪声	83	决策:把待识别样本归类到类中

正如 3.1 节所述，简单模式的句法结构一般为主谓宾结构，较为单一. 其中，特征谓词作谓语，往往位于句子的中心位置，而前置概念与后置概念分别作为特征谓词的主语(或宾语)和宾语(或主语)，分别位于特征谓词的两侧；同时，特征谓词一般为动词，前置概念与后置概念则多为名词. 因此，可以结合词位置与词性特征，构建演化关系抽取模板，利用模板匹配的方法进行简单模式的演化关系抽取. 具体的模板匹配模型如图 8 所示. 对于一个简单句 S，以分词后的词项为单位，词项个数 m 为句子长度，则特征谓词位于句子的 $\frac{m}{2}\pm i\left(i=0,1,2,\cdots,\frac{m}{2}-1\right)$ 位置且词性为动词(v)，前置概念与后置概念分别位于句子的 $\left(\frac{m}{2}\pm i\right)\pm j\left(i=0,1,2,\cdots,\frac{m}{2}-1,j=1,2,\cdots,\frac{m}{2}-1\right)$ 位置且词性为名词(n). 句子 S 中能够匹配这个模板的词则是演化关系主体(前置概念、特征谓词、后置概念).

1	*	*
2	*	*
⋮	⋮	⋮
$(m/2\pm i)-j$	$\#concept_1$	n
$m/2\pm i$	特征谓词	v
$(m/2\pm i)+j$	$\#concept_2$	n
⋮	⋮	⋮
m	。	wp

图 8 模板匹配模型

为了有效地进行基于模板匹配的演化关系抽取，分别对特征谓词与概念(前置概念、后置概念)设置了滑动窗口($i=0,1,2,3$；$j=1,2,3$)进行模板匹配的实验. 基于模板匹配的演化关系抽取结果如表 8 所示.

从表 8 的实验结果中可以发现，基于模板匹配的演化关系抽取的实验性能并不理想，当特征谓词在句子的中心位置或距中心位置距离为 1 的位置，概念词在相距特征谓词距离为 3 的位置时，基于

表 8　模板匹配的演化关系抽取结果

滑动窗口 i	滑动窗口 j	P/%	R/%	F/%
0	1	28.57	9.52	14.29
	2	44.44	19.05	26.67
	3	75.00	21.43	33.33
1	1	28.57	23.81	25.97
	2	37.04	23.81	28.99
	3	63.89	54.76	58.97
2	1	40.91	21.43	28.13
	2	42.86	28.57	34.29
	3	47.83	26.19	33.85
3	1	24.39	23.81	24.10
	2	29.03	21.43	24.66
	3	18.75	7.14	10.34

模板匹配的演化关系抽取结果相对较好：例如当 $i=0, j=3$ 时准确率能达到 75%，但是召回率偏低；而当 $i=1, j=3$ 时，召回率能达到 54.76%，而准确率仅有 63.89%. 同时，仅仅针对简单模式采用模板匹配的方法无法有效地应用到复杂模式. 此外，根据模板匹配获得的演化关系主体并不能区分出前置概念与后置概念，从而也就无法获取领域知识间的前序与后续的逻辑关系.

因此，本文实验不考虑利用机器学习进行演化关系模式分类，而是根据不同演化关系模式下具有的不同演化关系推理方法，统一进行演化关系抽取.

5.3　演化关系抽取结果

在训练条件随机场(CRF)模型时，本文对不同的特征集合组合(如表 9 所示)进行了训练，以挑选最合适的特征函数集合.

表 9　特征集的选择

特征集	包含特征	特征集	包含特征
F1	词 id	F7	F1+F2+F3
F2	POS 类型	F8	F1+F2+F4
F3	命名实体类别	F9	F1+F2+F5
F4	依存句法	F10	F3+F4+F5
F5	语义角色标注	F11	F1+F2+F3+F4+F5
F6	F1+F2		

针对不同特征集，条件随机场模型(CRF)的表现效果如表 10 所示，从实验结果来看，单特征的实验中效果较好的是依存句法与语义角色标注，因为这两个是基于句子层面而不是基于词的角度，而且在两者组合起来构成的特征集进行实验，效果也足够好；但当特征集选取词 id、POS 类型、命名实体类别、依存句法和语义角色标注时，不仅考虑了句子层面同时也考虑到单个词的因素，使得条件随机场(CRF)模型的表现结果最好.

表 10　实验结果

	P/%	F/%	R/%
F1	31.92	26.69	29.07
F2	40.58	31.04	35.17
F3	41.47	36.18	38.64
F4	70.34	57.73	63.41
F5	68.43	60.72	64.34
F6	44.86	38.67	41.54
F7	69.99	62.77	66.18
F8	72.58	64.38	68.23
F9	71.93	66.20	68.95
F10	75.40	70.02	72.61
F11	**80.66**	**77.93**	**79.27**

为了证明条件随机场模型对演化关系抽取的有效性，本文选择了隐马尔可夫模型(HMM)与最大熵马尔可夫模型(MEMM)在上述的数据集上进行了对比实验. 具体实验结果如表 11 所示.

表 11　CRF、HMM、MEMM 对比实验结果

	P/%	R/%	F/%
CRF(F11)	**80.66**	**77.93**	**79.27**
HMM(F11)	61.23	52.61	56.59
MEMM(F11)	77.97	73.63	75.74

从表 11 的实验结果可以看出：相对于隐马尔可夫模型(HMM)和最大熵马尔可夫模型(MEMM)，采用条件随机场(CRF)模型进行领域知识演化关系抽取能取得较好的准确率，实验性能更优. 其中，HMM 模型是将多特征合并构成新的单一特征作为 HMM 的特征进行训练，容易造成个别特征信息的丢失，导致实验效果不太理想；MEMM 模型并未将多特征合并构成新的单一特征，减少了特征稀疏的可能性，使得实验结果得到提升，但是 MEMM 只在局部做归一化，容易出现标记偏置；而在 CRF 模型中，综合了数据中的上下文信息，统计的是全局概率，在做归一化时，考虑了数据在全局的分布，而不是仅仅是在局部归一化，从而使得其在 3 种模型中表现最优.

同时对数据集中的原始实验数据进行分析可以发现影响实验结果的因素有以下两点：

(1) 使用中文自然语言解析文本时，可能会将一个完整的概念实体拆分，例如“支持向量机的理论基础来自于统计学习理论”①中，分词系统会将“支持向量机”、“统计学习”分别拆分成“支持”“向”“量机”、“统计”“学习”，分词效果不太理想. 这样就会影响后续的命名实体识别、依存句法分析与语义角色标注的效果，从而影响最终的实验结果. 在后续

① http://zh.wikipedia.org/wiki/统计学习理论

研究中可尝试使用自定义机器学习领域术语词典来减少因分词不当造成的影响.

(2)本文只针对中文维基百科的数据进行处理,而在数据集中,有些词条是包含英文的,例如"为了『教导』感知机识别图像,Rosenblatt 在 Hebb 学习法则的基础上,发展了一种迭代、试错、类似于人类学习过程的学习算法-感知机学习"①中,在使用中文解析工具进行数据处理之前,就已对数据进行了一次预处理,剔除数据中的非中文内容以及无关符号.由于剔除了数据中的非中文内容,因此会不可避免地对语句的语法结构分析造成影响.

同时,为了准确评估本文算法对于训练样本的依赖性,我们对简单模式和复杂模式的样本分别进行了实验评测和分析.首先将标记样本进行人工分类,筛选出简单模式样本和复杂模式样本,然后分别对两类模式的样本采用 5 倍交叉验证,按 4∶1 进行切分(即 4 份作为训练集,1 份作为测试集)运用 CRF 进行训练与测试,实验结果如表 12 所示.

表 12 简单模式、复杂模式实验结果

模式类别	数量	P/%	R/%	F/%
简单模式	221	82.51	72.45	77.15
复杂模式	623	90.74	85.92	88.26

由于过滤了实验数据中的噪声数据,并单独对简单模式与复杂模式进行演化关系的抽取,两种模式下的实验效果较表 11 得到显著提升;但是从实验结果上可以发现 CRF 在复杂模式样本上的实验性能更优,这一结论与我们预期的"简单模式更便于抽取"的设想相驳,我们发现产生该现象的原因与两种样本数据的规模有一定的关系,在公开出版的中文表达中,只有主谓宾的语法结构是比较少见的,常见的表达通常是在基本的主谓宾语法结构上添加其它修饰性的语法结构,将一个简单句扩展成复杂句.如"欧拉公式发展出了拓扑学"是一个简单句,但在实际的表述中变成了"欧拉提出的关于凸多边形顶点数、棱数及面数之间的关系的欧拉公式此后又被柯西等人进一步研究推广,发展出了拓扑学,成为了它的起源".因此,数据集中的简单模式样本相对复杂模式样本规模较小,使得简单模式的模型训练不足,影响了抽取算法的实验性能.

此外,我们从实验结果中统计了抽取出的正序演化关系与逆序演化关系的数据分布(如表 13 所示),可以发现正序演化关系与逆序演化关系的抽取结果数量相当,说明中文科普类文本有别于英文科普文本的表达习惯,没有一味侧重于使用被动方式(逆序)来表达语义关系.这一特性对于关系抽取与语言表达的关联研究将产生一定的积极作用.

表 13 演化关系类别统计

类别	占测试集比例/%	实例	抽取结果
正序关系	39	在现实中,三定律成为机械伦理学的基础	(三定律,*evolution*,机械伦理学)
逆序关系	44	公理化集合论起步于类型论	(类型论,*evolution*,公理化集合论)
噪声	17	监测:从图像中发现特定的情况内容	

简单模式与复杂模式中分别包含正序关系与逆序关系,本文在简单模式与复杂模式的演化关系抽取基础上,针对正序关系与逆序关系也分别进行了实验评测与分析(采用 5 倍交叉验证,按 4∶1 切分数据).实验结果如图 9 所示(其中括号内数字表示该类关系在数据集中的数量).

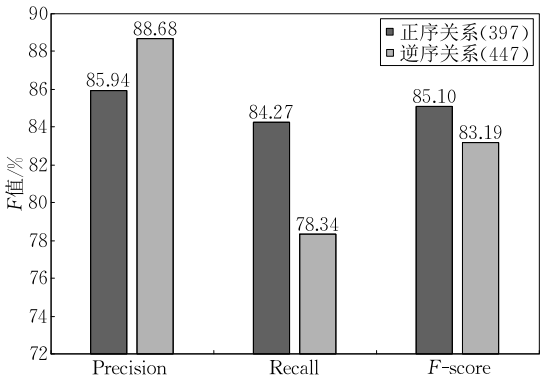

图 9 正序关系、逆序关系实验结果

图中数据表明本文算法对于正序关系与逆序关系都具有较好的抽取性能(F 值达到 83%以上),但从对比数据来看,逆序关系的召回率偏低,观察逆序关系数据发现,逆序关系数据中含有简单模式的数据较多,影响了 CRF 模型训练性能.如"深度学习的基础是分散表示"属于简单模式,但同时包含了逆序关系.

从实验结果中抽取出部分演化关系三元组,通过观察发现,可以对关系三元组中的概念实体构建领域知识演化关系图(如图 10 所示).通过知识演化关系图能够直观地发现知识之间的前后关联,对学习者学习和理解领域知识,梳理领域知识的前序和后续逻辑关系具有重要意义.同时,我们根据维基百

① http://zh.wikipedia.org/wiki/感知器

科中各个知识(概念)词条给出的该知识理论的诞生时间对部分领域知识进行排序,得到知识的时间序列图(如图 11 所示,部分知识因查找不到确切的时间,故未在图中显示).观察图 10 与图 11 可以发现具有演化关系的知识在演化序列和时间序列上也保持一致的先后顺序,即后置概念的出现是以前置概念的出现为前提,后置概念是以前置概念为基础.

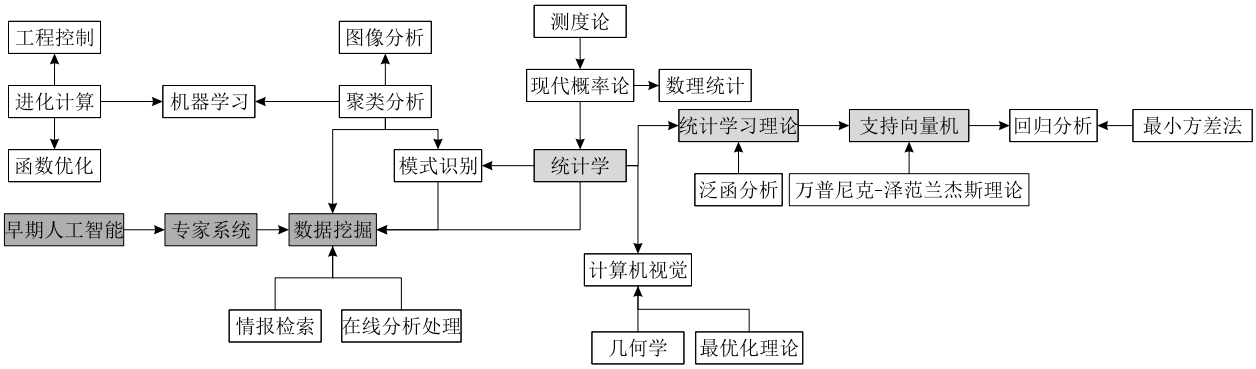

图 10　领域知识演化关系

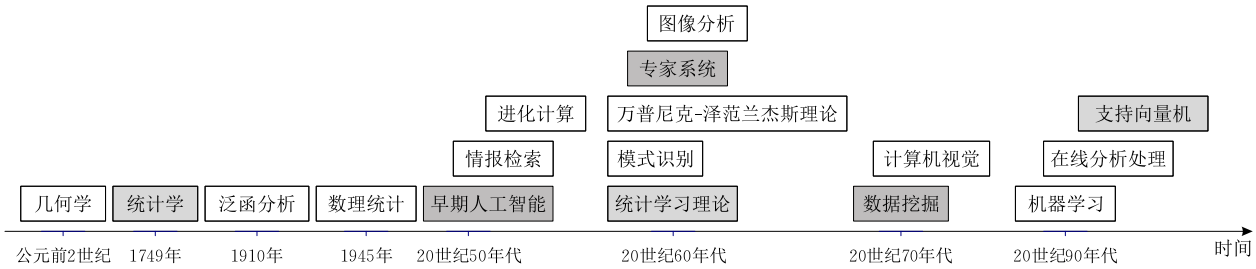

图 11　知识时间序列

经过 CRF 模型的关系抽取算法得到的演化关系三元组$\{\varepsilon_1,R,\varepsilon_2\}$,可以构成以 ε_1(前置概念)与 ε_2(后置概念)为点,R(演化关系)为有向边的有向图(如图 12 所示),表示演化关系由前置概念指向后置概念.

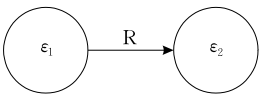

图 12　演化关系三元组有向图

因此,对于抽取出的每一个演化关系三元组$\{\varepsilon_1,R,\varepsilon_2\}$,均可以构成一个演化关系三元组有向图.而在一个演化关系三元组有向图中,其前置概念可能是另一个演化关系三元组有向图中的后置概念(同样的,其后置概念可能是另一个演化关系三元组有向图中的前置概念).因此,根据领域知识中不同概念的相互关系,将若干个演化关系三元组有向图首尾相连,就可以构成一个该领域的领域知识图谱.本文选取抽取的部分演化关系三元组,利用可视化工具 Echats[①] 构造了一个局部的"机器学习"领域知识图谱(如图 13 所示).不难看出,从构造的领域知识图谱能够充分地了解机器学习相关知识(概念)的前后发展与关联,为人们学习和理解相关理论提供准确、全面的前继和后继知识,对于建立和完善学科知识具有重要的指导意义.

同时从图 13 中可以发现,有些概念作为多数概念的后置概念,则其入度较多;同样的,有些概念作为多数概念的前置概念,则其出度较多.因此这些概念在知识图谱中起到了连接其它概念的重要作用.根据每个概念的入度和出度,我们给出该领域的部分核心概念和基础概念的领域知识图谱(如图 14 所示).例如,图 14 中"机器学习"的入度较多,则认为其他概念对该概念的贡献较大,因此"机器学习"可以作为该领域的核心概念(知识);同理,出度较多的概念(例如"人工智能")则可认为是该领域的基础知识,由它可演绎或推理得到该领域新的知识.由此,通过研究这些核心概念与基础概念可构建相关课程的重点、难点知识体系,对课程建设、知识工程等相关工作具有重要的研究意义.

① http://echarts.baidu.com/echarts2/index.html

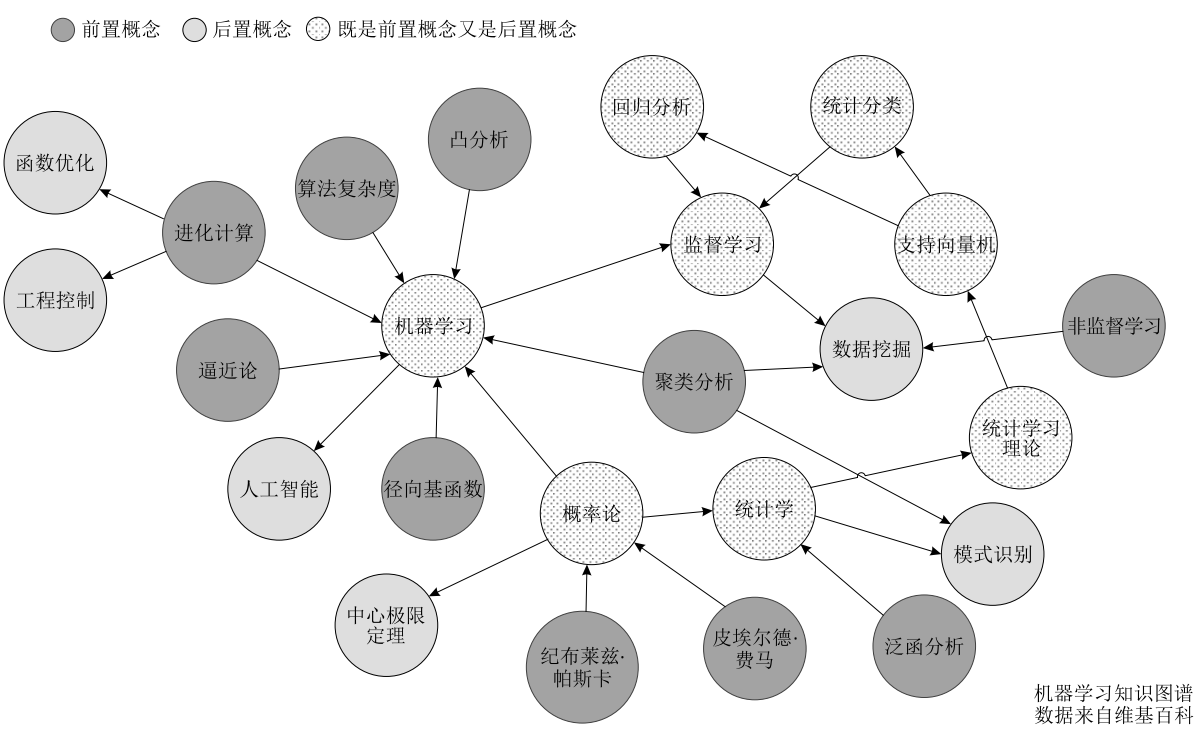

图 13 局部的机器学习领域知识图谱

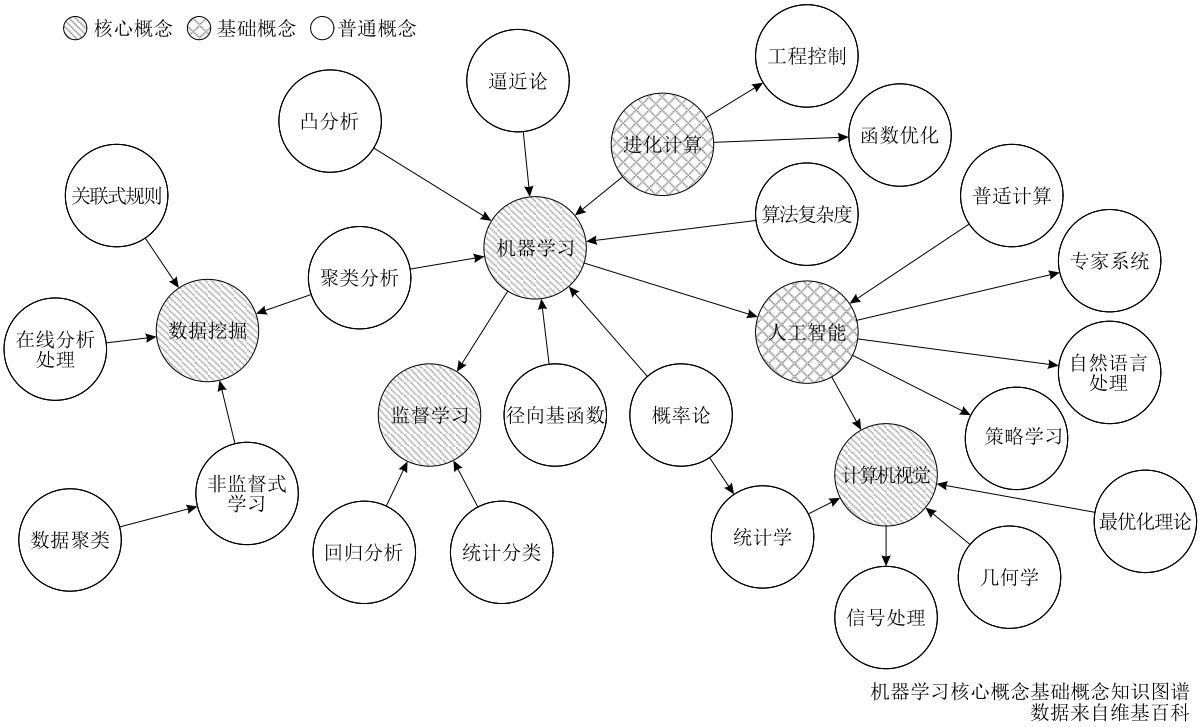

图 14 机器学习领域核心概念/基础概念知识图谱

6 结束语

互联网下的同一领域中不同知识概念间的演化关系对于用户学习和理解领域知识,梳理领域知识的前序和后续逻辑关系具有重要意义,但是网络数据的多样和无序使用户难以准确有序地获取领域知识间的演化关系.针对这一问题,本文提出一种基于条件随机场的面向中文维基百科领域知识的演化关系抽取方法,利用语法分析特征,构建演化关系推理模型.以词性、命名实体类别、依存句法结构以及语义角色等语言特征作为模式属性,结合序列标注,利

用 CRF 模型有效识别机器学习领域中概念间的演化关系.通过对比实验,验证了本文方法的优越性和有效性.在以后的研究工作中,考虑到依据演化推理模型进行人工标注的繁琐性,可以将演化推理模型进一步完善,利用基于规则或模式的方法进行演化关系的推理,最后运用半监督或无监督的方法进行演化关系的抽取.同时,在后续的研究工作中将对多个不同类型的百科数据集进行实验评测,进一步检验本文方法的普适性.

致　谢　审稿老师给出了宝贵的评语和修改意见,在此表示感谢!

参　考　文　献

[1] Wang Ping. Domain Knowledge Mining in Network Environments [Ph. D. dissertation]. East China Normal University, Shanghai, 2010(in Chinese)
(王萍. 网络环境下的领域知识挖掘[博士学位论文]. 华东师范大学, 上海, 2010)

[2] Chen Chi, Wang Yu-Peng, Li Chao, et al. Research and application of big data in the field of online education. Journal of Computer Research and Development, 2014, 51(Supplement): 67-74(in Chinese)
(陈池, 王宇鹏, 李超等. 面向在线教育领域的大数据研究及应用. 计算机研究与发展, 2014, 51(增刊): 67-74)

[3] Wang Yuan-Zhuo, Jia Yan-Tao, Liu Da-Wei, et al. Open web knowledge aided information search and data mining. Journal of Computer Research and Development, 2015, 52(2): 456-474(in Chinese)
(王元卓, 贾岩涛, 刘大伟等. 基于开放网络知识的信息检索与数据挖掘. 计算机研究与发展, 2015, 52(2): 456-474)

[4] Hu Bao-Shun, Wang Da-Ling, Yu Ge, et al. An answer extraction algorithm based on syntax structure feature parsing and classification. Chinese Journal of Computers, 2008, 31(4): 662-676(in Chinese)
(胡宝顺, 王大玲, 于戈等. 基于句法结构特征分析及分类技术的答案提取算法. 计算机学报, 2008, 31(4): 662-676)

[5] Zhang Hai-Su, Ma Da-Ming, Deng Zhi-Long. Semantic knowledge bases construction based on Wikipedia. Application Research of Computers, 2011, 28(8): 2807-2811(in Chinese)
(张海粟, 马大明, 邓智龙. 基于维基百科的语义知识库及其构建方法研究. 计算机应用研究, 2011, 28(8): 2807-2811)

[6] Wang Juan, Cao Shu-Jin, Jiang Ling-Min, et al. Research on semantic relatedness of domain-specific concepts based on Chinese Wikipedia. Library and Information Service, 2014, 58(23): 136-142(in Chinese)
(王娟, 曹树金, 姜灵敏等. 基于中文维基百科的领域概念相关性研究. 图书情报工作, 2014, 58(23): 136-142)

[7] Hua Bo-Lin. Architecture of knowledge extraction based on NLP. New Technology of Library and Information Service, 2007, 10: 38-41(in Chinese)
(化柏林. 基于 NLP 的知识抽取系统框架研究. 现代图书情报技术, 2007, 10: 38-41)

[8] Liu Li, Xiao Ying-Yuan. A verb dependency set based domain concept clustering method. Journal of Harbin Engineering University, 2015, 36(7): 1-5(in Chinese)
(刘里, 肖迎元. 基于动词依存集的领域概念聚类方法. 哈尔滨工程大学学报, 2015, 36(7): 1-5)

[9] Hu Yan. Research on Specialty Knowledge Retrieval Method Based on Web Information Extraction [Ph. D. dissertation]. Wuhan University of Technology, Wuhan, 2007(in Chinese)
(胡燕. 基于 Web 信息抽取的专业知识获取方法研究[博士学位论文]. 武汉理工大学, 武汉, 2007)

[10] An Xin-Ying, Leng Fu-Hai. Study on disjoint literature-based knowledge discovery. Journal of the China Society for Scientific and Technical Information, 2006, 25(1): 87-93(in Chinese)
(安新颖, 冷伏海. 基于非相关文献的知识发现原理研究. 情报学报, 2006, 25(1): 87-93)

[11] Hong Jun. Deep Learning Based Domain Concept Extraction Algorithm [M. S. dissertation]. East China Normal University, Shanghai, 2014(in Chinese)
(洪俊. 基于 Deep Learning 的领域概念抽取方法研究 [硕士学位论文]. 华东师范大学, 上海, 2014)

[12] Dong Li-Li, Li Huan, Zhang Xiang, et al. Method for automatic extraction of chinese domain concepts. Computer Engineering and Applications, 2014, 50(6): 127-131(in Chinese)
(董丽丽, 李欢, 张翔等. 一种中文领域概念词自动提取方法研究. 计算机工程与应用, 2014, 50(6): 127-131)

[13] Tsymbal A. The problem of concept drift: Definitions and related work. Ireland, TCD-CS-2004-15 (2004) Learning of Variable Rate Concept Drift 151, 2004

[14] Gama J, Medas P, Castillo G, et al. Learning with drift detection//Proceedings of the 17th Brazilian Symposium on Artificial Intelligence. Berlin, Germany, 2004: 286-295

[15] Baena-Garcıa M, del Campo-Ávila J, Fidalgo R, et al. Early drift detection method//Proceedings of the 4th International Workshop on Knowledge Discovery from Data Streams. Berlin, Germany, 2006: 77-86

[16] Nishida K, Yamauchi K. Detecting concept drift using statistical testing//Proceedings of the 10th International Conference on Discovery Science. Berlin Heidelberg, Germany, 2007: 264-269

[17] Wang H, Fan W, Yu P S, et al. Mining concept-drifting data streams using ensemble classifiers//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2003: 226-235

[18] Han Meng, Wang Zhi-Hai, Yuan Ji-Dong. Efficient method for mining closed frequent patterns from data streams based on time decay model. Chinese Journal of Computers, 2015, 38(7): 1473-1483(in Chinese)
(韩萌, 王志海, 原继东. 一种基于时间衰减模型的数据流闭合模式挖掘方法. 计算机学报, 2015, 38(7): 1473-1483)

[19] Radhakrishnan P, Gupta M, Varma V. Modeling the evolution of product entities//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Gold Coast, Australia, 2014: 923-926

[20] Liu Y, Jin P, Yue L. Extracting position relations from the web//Proceedings of the 11th International Workshop on Web Information and Data Management. Hong Kong, China, 2009: 59-62

[21] Zhang L, Li L, Li T, et al. Patentline: Analyzing technology evolution on multi-view patent graphs//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Queensland, Australia, 2014: 1095-1098

[22] Maslennikov M, Chua T S. Combining relations for information extraction from free text. ACM Transactions on Information Systems (TOIS), 2010, 28(3): 1-35

[23] Chiang Y H, Doan A H, Naughton J F. Modeling entity evolution for temporal record matching//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird, USA, 2014: 1175-1186

[24] Zhang Wei-Ru, Sun Le, Han Xian-Pei. A entity relation extraction method based on Wikipedia and pattern clustering. Journal of Chinese Information Processing, 2012, 26(2): 75-83(in Chinese)
(张苇如, 孙乐, 韩先培. 基于维基百科和模式聚类的实体关系抽取方法. 中文信息学报, 2012, 26(2): 75-83)

[25] Yu Xiaofeng, Lam Wai. An integrated probabilistic and logic approach to encyclopedia relation extraction with multiple features//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, England, 2008: 1065-1072

[26] Tang Qing, Lv Xue-Qiang, Li Zhuo. Research on domain ontology concept hyponymy relation extraction. Micro Electronics & Computer, 2014, 31(6): 68-71(in Chinese)
(汤青, 吕学强, 李卓. 本体概念间上下位关系抽取研究. 微电子学与计算机, 2014, 31(6): 68-71)

[27] Caraballo S A. Automatic construction of a hypernym-labeled noun hierarchy from text//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Providence, USA, 1999: 120-126

[28] Huang Yi, Wang Qing-Lin, Liu Yu. An acquisition method of domain-specific terminological hyponymy based on CRF. Journal of Central South University(Science and Technology), 2013, 44(2): 355-359(in Chinese)
(黄毅, 王庆林, 刘禹. 一种基于条件随机场的领域术语上下位关系获取方法. 中南大学学报(自然科学版), 2013, 44(2): 355-359)

[29] Jia Li-Li. The Study of Concept-to-Concept Relations in Ontology Construction[M. S. dissertation]. Chinese Academy of Agricultural Sciences, Beijing, 2007(in Chinese)
(贾黎莉. Ontology 构建中概念间关系的研究[硕士学位论文]. 中国农业科学院, 北京, 2007)

[30] Zhu Li-Jun, Tao Lan, Huang Chi. Semantic web: Concept, approach and application. Computer Engineering and Applications, 2004, 40(3): 79-83(in Chinese)
(朱礼军, 陶兰, 黄赤. 语义万维网的概念、方法及应用. 计算机工程与应用, 2004, 40(3): 79-83)

[31] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). San Francisco, USA, 2001: 282-289

[32] Zhang Qi, Jin Pei-Quan, Yue Li-Hua. CRF based dynamic relations extraction from web. Journal of University of Science and Technology of China, 2010, 40(11): 1197-1202 (in Chinese)
(张奇, 金培权, 岳丽华. 基于 CRF 的网页动态关系抽取研究. 中国科学技术大学学报, 2010, 40(11): 1197-1202)

GAO Jun-Ping, born in 1991, M. S. candidate. His research interest is information extraction.

ZHANG Hui, born in 1972, Ph. D., professor. His research interests include text mining, knowledge engineering.

ZHAO Xu-Jian, born in 1984, Ph.D., associate professor. His research interests include Chinese information processing, Web information retrieval.

YANG Chun-Ming, born in 1980, M. S., associate professor. His research interests include text mining, knowledge engineering.

LI Bo, born in 1977, Ph. D., lecturer. His research interests include information filtering, information security.

Background

The research on evolutionary relation extraction, as a kind of relation extraction, has a vital important role in Online Learning, Knowledge Graph Construction, etc., which can help users to learn and understand the domain knowledge as well as sorting out the successive logic relationship between two different concepts in the same field of knowledge from Internet. However, the diversity and disorder of the network data make it difficult for users to obtain relations among domain knowledge accurately and orderly.

This paper focuses on this issue, and surveys the relevant research work that aimed to address the problem. And this paper proposes an evolutionary relation extraction method for domain knowledge in Chinese Wikipedia. Firstly, the authors construct a reasoning model for evolutionary relation using different syntax features, and then detect the evolutionary relation for domain knowledge taking advantage of a sentence level relation extraction algorithm. Experiments show that the proposed method can extract evolutionary relation in domain knowledge effectively.

This paper is supported in part by the Scientific Research Funds in Sichuan Province Department of Education under Grant No. 14ZB0113, by the Doctoral Research Fund of SWUST under Grant No. 12zx7116, by the Innovation Project of next generation Internet technology of Saier under Grant No. NGII20150510.

Researchers started the research on information extraction and text mining in 2008, and have achieved some promising results involved in relation extraction, topic model and topic evolution detection. More detailed can be found on researchers' publications.