DOI:10.13530/j.cnki.jlis.160003

学术论文中方法知识元的类型与描述规则研究*

化柏林

摘　要　学术论文中有很多方法知识元的描述,如何把这些方法知识元抽取出来,形成结构化的方法知识库,是细粒度知识组织的重要研究内容之一。本文通过对大量的文献进行内容分析,把方法知识元归结为方法定义知识元、方法关系知识元、方法特点知识元、方法流程知识元和方法功能知识元五种类型。对论文中关于方法描述的句子进行抽取,通过过滤句子中的领域关键词形成句子描述结构,在此基础上经过人工审核与合并归类,形成方法知识元的描述规则,为后续的方法知识元抽取提供支撑。图 3。表 7。参考文献 17。

关键词　学术论文　方法知识元　知识挖掘　模式识别

分类号　G302

Types and Description Rules of Knowledge Elements About Method in Academic Papers

HUA Bolin

ABSTRACT

In academic papers there are many knowledge elements about method. In order to construct a structural method knowledge base, we need to extract these elements. These elements form a key data source for a method system. The extraction of knowledge elements about methods is an important research topic in deepening knowledge organization research in the direction of finer granularity. With a knowledge base on methods, not only can we draw method tree diagram and development map, but also it can be embedded into a decision support system or an intelligent system in which it acts as a source of method selection. This will facilitate the method usage standard and development.

We select 17 LIS core journals from CSSCI and meta data of articles in these journals are downloaded from CNKI, WanFang Data and CQVip. After data fusion and cleansing, we do a statistical analysis on keywords and get a keyword term list which includes 63 203 keywords. 1 302 of these keywords are identified as method terms. Thus we have a method term list. Then method term list is used to do a full text recognition in all papers in Journal of The China Society For Scientific and Technical Information, 2012. Among all the 18 686 sentences, 2 707 are recognized as a description of a method. We do word segmentation on all these sentences. The word lists we use for word segmentation are a Chinese dictionary and a LIS domain keyword list.

* 本刊“青年学术论坛”特约稿(Special contribution for the Youth Academic Forum sponsored by this Journal)

本文系中国博士后科学基金项目“学术论文中方法知识元的挖掘与发现研究”(编号:2014M560857)的研究成果之一。(This article is an outcome of the project “Mining and Discovery of Knowledge Cell from Academic Papers” (No. 2014M560857) supported by China Postdoctoral Science Foundation.)

通信作者:化柏林,Email: huabolin@pku.edu.cn,ORCID: 0000-0001-9248-6455(Correspondence should be addressed to HUA Bolin, Email: huabolin@pku.edu.cn,ORCID: 0000-0001-9248-6455)

Both the keyword list and a domain subject term list are used for filtering domain words in these sentences. Every sentence forms a linear structure, which is a syntactic structure on method knowledge elements. For example, we have "the method of… is a ", "the method of … has certain disadvantages" , "… method has been adopted to solve …". After we have a list of such structures on hands, we do a manual inspection on the type of the method knowledge element. At last, we end up with a method knowledge elements rule base.

We have identified 5 types of method knowledge elements, i.e. method definition knowledge element, method relation knowledge element, method feature knowledge element, method process knowledge element and method functionality knowledge element. Method relation knowledge element includes static spatial relation and dynamic evolution temporal relation. Method feature knowledge element includes both the strength and weakness of a method. Some of the feature descriptions are discussions on a single method while others are comparisons on a one-to-one basis or a one-to-many basis.

The result shows that different types of method knowledge elements have different features and rule descriptions. Method definition and functionality knowledge elements are comparatively simple, so are the sentence rules. Most method feature knowledge elements use comparative sentence in the comparison with another method or several other methods. Their syntactic structures are comparatively complex. There are not so many static relation descriptions between methods. But the descriptions of the rule are complex. Meanwhile, there are many dynamic relation descriptions; however, the descriptions of the rule are not complex. Compared with a single sentence, a sentence group or a paragraph is more suitable in the description of method process knowledge element. It is difficult for us to construct rules for such knowledge elements.

The study presented in this paper has its weakness which can be further studied. On one hand, there is no test dataset on method extraction, so it is difficult to evaluate our result. On the other hand, the rules we constructed cannot cover all the situations and the applicability of the study. In the future we will expand the scale of the raw corpus and construct a method knowledge element test dataset which can be used to evaluate the performance of method knowledge element extraction. 3 figs. 7 tabs. 17 refs.

KEY WORDS

Academic papers. Knowledge element of method. Knowledge mining. Pattern recognition.

0 引言

作为知识的重要载体,学术论文是科研成果的重要体现与科研创新的结晶。在学术论文中,方法的描述是科学知识的一种重要类型。随着学术论文数量的急速增长,仅靠人力已经难以胜任方法的监测分析与方法库的构建工作。现在的数字环境越来越多地依赖信息技术,只有充分利用技术手段,通过知识抽取与挖掘从大规模文献中获取有价值的信息与知识,才能快速有效地建成较为全面的方法知识库。而要实现自动或半自动地从论文中抽取与挖掘方法知识元,需要对方法知识元的类型与描述规则进行深入分析与研究。

方法既包括调查问卷、专家访谈、案例分析等一般科学研究方法,也包括计量分析、聚类分析、关联分析、共现分析、多目标决策分析等问题解决方法。这些方法经常出现在学术论文中,因此从学术论文中抽取方法术语,可以为方法体系的构建提供关键的数据源,形成方法知识库。方法知识库既可以嵌入到决策支持系统或专家智能系统,提供方法的选择与支撑,也可以在此基础上绘制方法谱系图和学科方法发展

地图,促进学科对方法的规范性使用与发展。因此,关于方法知识元的内容抽取与挖掘具有重要的学术研究价值与实践应用意义。

知识挖掘主要有统计或规则两条技术路线。在以词为处理单元的文本挖掘中,统计学习的方法盛行于规则方法,但在以句子为单元的文本挖掘研究中,由于句子的复杂度等原因,统计学习方法难以适应,于是不少学者采取规则与模式识别的方法。学术论文中方法知识元的描述多以句子为单位,这种情况比较适合规则与模式识别的方法。方法知识元包括哪些类型,以及如何构建方法知识元的描述规则,就成为从论文中抽取方法知识元的关键。

1 研究综述

关于知识元以及规则抽取,国内外已有一些相关研究,这些研究集中在知识元理论与述评、知识元表示与建模、知识元抽取与实现以及规则的自动抽取方面。

(1)知识元理论与述评研究

温有奎等认为,知识元语义链接理论将代表未来知识发现模式,从知识元语义链接的角度详细论述文献知识元间隐含关联的潜在知识发现的基本概念、方法和技术[1]。高继平等对知识元的定义、知识元的计量指标、知识元的研究项目、知识元在不同学科领域的作用及研究现状进行述评[2]。姜永常认为知识组织应以知识元为基元,以知识元链接为枢纽来构建知识组织的神经系统[3]。文庭孝等对中文文本知识元构建的意义及困难进行分析,认为中文分词会成为知识元抽取的技术瓶颈[4]。分词是中文文本信息处理的基础工作,但并不是影响知识元抽取的关键困难,英文知识元抽取也存在同样的问题,在确定知识元的时候,需要把多个英文词组合起来形成一个知识元概念,如"Knowledge Management"作为一个单元来表示知识元,如果只用 Knowledge 或 Management 来表示知识元是没有意义的。

(2)知识元表示与建模研究

在知识元表示与建模方面,王宇等在期刊文献知识元库的基础上提出了一种基于六元组知识元(编号、导航、来源、类型、特征词、内容)的期刊文献知识仓库的构建方法,设计了从知识元库到知识仓库的知识抽取的方法[5]。姜永常等基于 Brooks 文献中的知识节点及 Swanson 文献间的隐性关联方法,提出一种基于知识元本体语义链接的知识网络构建方法和实现模型[6]。仲秋雁等通过抽取情景共性要素及要素关系提出情景元模型,在此基础上提出具体领域的基于知识元的情景概念模型,而决策人员面对的具体情景则是对情景概念模型的实例化[7]。

(3)知识元抽取与实现研究

在知识元抽取与实现方面,温有奎等通过对科技论文知识创新生产、知识增值管理、知识集成利用等方面的探讨,分析文本创新点的表现形式,对创新点的挖掘做了试验,结果表明基于创新点的知识元挖掘是文本知识挖掘的一种有效方法[8]。冷伏海等综合运用语义标注、规则抽取以及正则表达技术,提出了一种混合语义信息抽取方法,从科技文献中抽取其主要研究方法、性能指标,既不破坏科技文献原有语义内容,又能以较为简单的方式展示科技文献的主要创新内容[9]。周宁等提出一种基于 XML 平台的知识元表示与抽取模型,将文档分解为许多段落,从段落中解析出相应的基本知识元,用结构约束、长度约束和内容约束来表示知识元,并通过结构解析、长度解析和内容解析三个步骤进行知识元的抽取[10]。朱丽萍等对背景知识、问题分析、文章所做工作等引言三要素进行分析研究,总结引言三要素的常用句型及特征,利用这些规则对引言三要素信息进行结构化抽取,将生物医学文献全文中的句子自动归类到引言、方法、结果与讨论中[11]。

(4)规则的自动抽取研究

德国多特蒙德大学用无监督学习算法的神经网络从事实型数据中抽取规则,然后把这些规则转成 PROLOG 规则[12]。谢孟军等提出一

种基于理论的规则自动抽取的设计方案[13]。孙晨等认为尽管神经网络已经在很广泛的领域得到应用，但由于训练好的神经网络中的知识不易于理解，可从神经网络中抽取规则来表示其中隐含的知识，以解决这一问题[14]。侯广坤等应用决策树归纳学习的优化原则，使得生成的决策树能最简洁、准确地描述从神经网络中学到的知识[15]。高阳等提出一种自适应的概率规划规则抽取算法，在强化学习获得的最优状态—动作对值函数基础上，通过 Beam Search 算法从值函数中抽取满足概率规划条件的规划知识[16]。

从这些研究可以看出，随着认知理论的不断发展以及自然语言处理能力的提高，对文献正文内容进行抽取与挖掘正得到逐步重视。这些研究主要集中在两个领域。

(1)在图书情报领域，学者们从理论方法、技术模型以及应用平台等角度展开了很好的探索与研究，并对学术定义、论文创新点等进行抽取研究，但专门针对论文里的方法进行内容抽取与挖掘的研究还不够充分，如何借助技术手段构建知识元描述规则的问题并没有很好地解决。

(2)在人工智能领域，在规则的自动抽取方面已有不少成果，但往往是智能学习或推理过程中的形式化规则，并不是从原始文献里进行知识元抽取的规则。

因此，本文以学术论文为研究对象，着力研究学术论文中方法知识元的类型与描述规则，对方法知识元的类型进行归纳总结，采用半自动的方法初步构建方法知识元的描述规则，为后续的知识抽取提供理论基础与资源支撑。

2 研究方法

2.1 研究的数据与素材

依据 CSSCI 选取 17 种图书情报领域核心期刊论文，分别从中国知网、万方数据以及重庆维普下载相关题录信息，对题录信息进行融合清洗以及汇总后，统计关键词，得到图书情报领域关键词表，有63 203条，从这些关键词中识别出方法术语，有1 302个，构建方法术语表[17]。利用方法术语表对《情报学报》2012 年全年的全文进行识别，从 18 686 个句子中共识别出2 707个关于方法描述的句子，对这些句子进行规则的识别与构建。

2.2 研究的流程与方法

首先读取每篇文章全文，运用方法术语识别含有方法的句子，然后利用中文词典和图书情报领域关键词库对这些句子进行分词，运用关键词库与领域主题词表对句子中的领域词进行过滤，形成句子的线性结构，即方法知识元的句型结构，例如，“……方法是一种……”“……方法存在……的缺点”“采用 ……方法对……进行……”，得到句式结构以后进行人工审核校对并判定方法知识元的类型，把这些关于方法描述的句式结构进行归类总结，形成方法知识元规则。构建过程如图 1 所示。

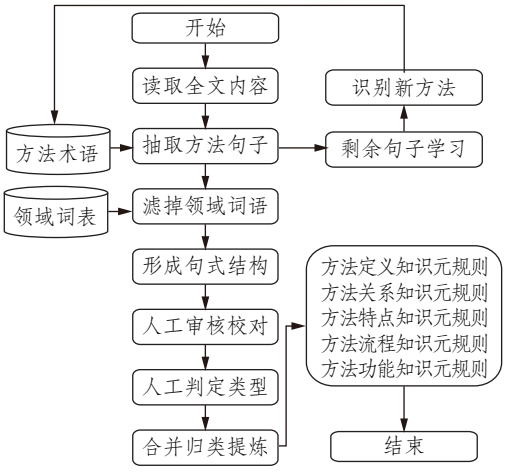

图 1 方法知识元规则构建流程

3 研究结果

如何刻画与描述方法知识元是一个关键问题，针对这个问题，本文提出方法知识元的五要素，即方法的定义、方法的特点、方法的关系、方法的流程以及方法的功能。由这五要素构成方

法知识元的五种类型，即方法定义知识元、方法关系知识元、方法特点知识元、方法流程知识元、方法功能知识元，如图2所示。

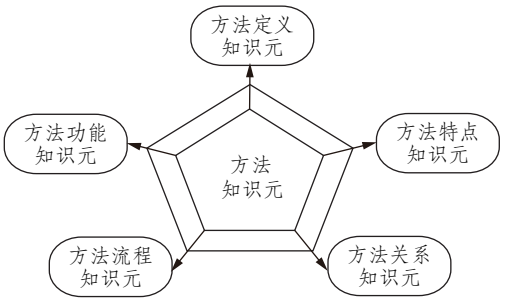

图2 方法知识元的五种类型

3.1 方法的定义描述

方法的定义是指对于某种方法的本质特征或关于方法术语概念内涵和外延确切而简要的说明。方法的定义描述通常具有以下规则“……方法是一种……的方法”“……方法+是|是指|指|的定义为|被定义为……”。利用规则对方法的定义描述进行识别，但有些句子即使符合这种规则，也有可能不是定义，称为伪定义句子，例如“局部分析方法是一种计算量小且不依赖于外部资源，但十分有效的查询扩展方法”是指方法的特点，而“人际竞争情报网络动态分析从本质上讲是一种网络分析方法”是指方法的类属。方法的定义描述规则及举例见表1。

表1 方法的定义规则及举例

规则	举例
……|法|方法+是一种……的方法。	内容分析方法是一种采用规范方法读取文本内容，将文本中信息有序、量化地表示出来的一种基于定量分析的定性研究方法。
……方法+是|是指|指|的定义为|被定义为|定义如下……	跨语言信息检索中，我们将翻译概率的计算方法定义如下：$P(f_i \mid e) = F(f_i)/(F(f_i)+F(\overline{f_i}))$ 在该公式中，$P(f_i \mid e)$表示f_i为检索词e的翻译的概率；$F(f_i)$为在一定的样本空间中，f_i为检索词e的翻译的频率；$F(\overline{f_i})$为在一定的样本空间中，检索词e的翻译不为f_i的频率，即检索词e翻译为其他译项的频率。

3.2 方法的关系描述

方法之间的关系描述是方法类属的体现，也是构建方法体系的基础。方法之间的关系从空间分布视角看包括上下位类的类属关系与同一层次的并列关系，不同的方法位于不同的层次和位置；从时间逻辑视角看方法之间的关系包括改进关系、继承关系、演进关系、替代关系等。

3.2.1 方法之间的静态关系

方法之间的静态关系主要描述方法的类别或属性关系，可以用“……方法+是|属于……一种”等规则来抽取，有些具体的方法列举包括数字序号型、汉字序号型。还有一种是带破折号的情况，如“本文从关联规则挖掘领域引入了一种新的共现聚类分析方法——最大频繁项集挖掘”。方法的类属描述规则及举例见表2。

方法类属描述的识别存在以下难点：伪关系的识别、上位类方法的缺省等。“……方法是一种……的方法”，这样的句式可能是描述方法之间的关系，也有可能是方法的定义。有些方法的列举找不到上位类方法，例如：“本文仅仅是探索性的研究，研究的结果还需要其他的方法来佐证，譬如用文献共被引分析、作者共被引分析等方法来验证和修正结论。”“借鉴社会网络的思路，作者共被引关系也可以进一步网络化，从而借助网络结构分析、凝聚度和中心性分析等方法可以对特定领域内作者的影响力情况进行深入的探索，以期对研究工作的推进和学科领域的发展提供一定的参考和帮助。”

表 2　方法的静态关系描述规则及举例

规则	举例
……方法+是|属于……一种	人际竞争情报网络动态分析从本质上讲是一种网络分析方法。 本文主要讨论术语的共现分析，又称共词分析，它属于内容分析方法的一种，不过本文提出的方法同样适用于其他研究对象的共现分析，如文献共引分析、文献作者共著分析等。
[根据|按照]……划分为|有……如|分别是|具体有……	聚类算法根据算法的不同划分为不同的类型，如基于划分的方法，基于层次的方法，基于密度的方法，基于网格的方法和基于模型的方法。 现有的专利本体构建方法有三种思路：分别是自顶向下法、自底向上法、中间扩展法。
……分为……等[几类]	根据消歧知识来源的不同，语义消歧可分为基于知识的方法和基于统计的方法，基于知识的方法又可细分为基于规则的方法和基于词典的方法。
方法+总结|包括|分为+以下|下面|如下+N+类|种|类型：	目前关于文本分类算法的研究很多，概括起来主要分为以下几类：①基于统计的方法，如朴素贝叶斯、KNN、类中心向量、支持向量机、最大熵等方法；②基于连接的方法，如人工神经网络；③基于规则的方法，如决策树等。
方法+如 M1，M2，M3[和|以及]Mn	传统的竞争情报分析方法，如 SWOT 分析方法、定标比超分析方法、关键成功因素分析方法、核心竞争力分析方法等。 自从最大频繁项集的概念被提出之后，研究者们提出了许多挖掘 MFI 的高效算法，如 MaxMiner、MAFIA、GenMax、Pincer Search、基于 Diffset 的方法以及 HBMFI 等。
……统称为|并列为……的几+类|大+方法	基于案例推理、基于规则推理、基于模型推理并列为“知识推理”的三大推理方法。

3.2.2　方法之间的动态关系

方法之间的动态关系包括改进关系、继承关系、演进关系、替代关系等，这些关系在具体论述中表现为“提出”“改进”等。创新地提出方法是指针对新的问题或基于新的数据或者面向新的需求，创新性地提出一种新的解决方法或方案。创新提出方法的规则比较简单，绝大多数创新描述都使用“提出”等特征词，也有使用“针对……设计了……”等句式。但有时候也存在噪声，如“提出一种……方法的改进”，虽然有“提出”，但实际上是改进。移植或改进方法是指把其他学科方法引入本学科，并对其适用性进行判断分析，在此基础上进行改进，或者直接改进本学科现有的方法。方法的动态关系描述规则及举例见表 3。

表 3　方法的动态关系描述规则及举例

规则	举例
提出了改进的……	提出了改进的两步式 K 核分析方法
提出……改进|修正	本文提出一种改进的近邻传播算法，使用该方法对 Web 用户进行聚类。 有学者提出了加权 PageRank 算法来改进传统的 PageRank 方法。
从……引入|借用|参考了……	本文从关联规则挖掘领域引入了一种新的共现聚类分析方法——最大频繁项集挖掘，它将传统共现分析法的三个阶段压缩为一个阶段，充分利用了可以利用的各种信息，克服了传统方法的缺陷。
将……引入到……以|来……	ListNet 和 ListMLE 将 Luce 模型引入到排序学习方法之中来表示文档序列并取得更好的排序效果。

3.3 方法的特点描述

方法的特点描述按照特点的褒贬分为优点描述、缺点描述与中性特点描述。按照描述的方法分为单纯论述型描述、对比论述型描述。对比论述型描述包括一对一比较型和一对多比较型两种类型。方法的特点描述类型见图3。

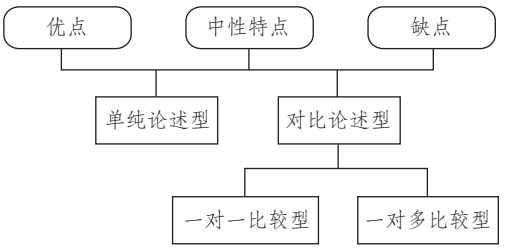

图3 方法的特点描述类型图

3.3.1 方法的优缺点描述

方法的缺点描述往往伴随着“无法”“问题”“缺点”“不足”等带有否定倾向的特征词。其规则与举例见表4。方法的优点描述往往不像缺点那样有明显的特征词,优点的描述往往通过具体的褒义词来确定(见表5)。对于优点的判别,可以使用褒义词表进行抽取。

表4 方法的缺点描述规则及举例

规则	举例
……不能|无法|……	但面对高维稀疏的数据,近邻传播算法往往不能得到很好的聚类结果,而且该方法不能产生指定类数的聚类。
……,[但|可是]存在|尚存|有……的缺点|不足	而该方法存在两个明显的缺点,其一是在训练统计翻译模型t(qi|w)时,需要大量的训练数据,而这在实际应用中很难获得。 这种方法的优点是容易实现,但它有两方面的缺点,一是某些常用的单个标签缺少语义的明确性,不能体现用户的个性化偏好,如标签“Web”等;二是用户在对具体对象进行标签时,同时使用了多个标签,它们之间存在着一定的语义联系,向量型的离散表示不能刻画标签之间的这种语义联系。
……极易+造成|带来……	现有的混合多准则决策问题的解决方法极易造成准则信息的丢失,同时,计算和转换的过程过于复杂。

表5 方法的优点描述规则及举例

规则	举例
……以……成为|为……的……	共现分析是基于定量化的数据分析,以其方法的简明性和分析结果的可靠性成为支撑信息内容分析研究过程的重要手段和工具。
对于……,……是一种……的方法	对于处理大规模的数据集,近邻传播算法是一种快速、有效的聚类方法。
……方法+综合了|兼顾了……的优点,……	采用信息熵的参数确定方法综合了计算法和实验法的优点,操作简单、方便,更能反映出参数与样本联系的本质特征,客观刻画了样本的重要性程度。
……方法+克服了|避免了……的问题|缺点,……	由于本文方法不需要先验知识,通过对数据集的全局进化优化,输出规模更小、结构简单便于理解的规则集,通过否定选择分类器进行规则匹配,克服了其仅适合样本稀疏或小样本空间的问题,在分类效果上优于决策树Assistant R方法,同时在分类效率上高于K-NN、贝叶斯方法。
与……相比,……方法[在……方面]+更具|好处是……	与传统聚类相比较,三维空间聚类法最大好处是直观、灵活,能够很好地揭示词义间内在联系。
……表明,……方法[在……方面]+更具……	实验表明,用SVM与修正的KNN组合算法进行专有名词抽取比单一SVM方法以及传统的SVM-KNN方法更具优越性,而且这种方法可以推广到其他非平衡分布样本的分类问题。

3.3.2　方法特点的比较句类型

方法特点知识元在句式上包括单纯论述型、一对一比较型以及一对多比较型，其规则与举例如表6所示。认识并揭示比较句的类型与规律，有助于更好抽取与挖掘方法特点知识元。单纯论述型是指没有比较的对象，单纯地对方法的特点进行描述。一对一比较型，其比较体往往是一个具体的方法术语，句子中会有明确的比较特征词，如“与……相比，……方法……”“相比……，方法……”。

表6　方法特点描述规则及举例

类型	规则	举例
单纯论述型	……方法……具有一定的……	该方法对军事领域的情报分析或是企业管理领域的决策信息处理均具有一定的适用性和参考意义。
	……方法……适用于……	本文提出的新的相似度测量方法概念明确，逻辑严谨，计算过程相对简单，特别适用于混合型多准则决策。
一对一比较型	与……相比，……方法……	与需要不断迭代的FlokRank方法相比，基于网络结构的推荐方法计算复杂性要小得多。 该算法可以提高多值属性关联规则挖掘效率，与Maqa算法相比，降低了时间复杂度，为处理海量数据、生成有用的知识和信息、进行知识管理及决策提供一种有效的方法。
	相比……，方法……	相比SVM方法，决策树方法能够学习析取规则表达式，这些规则可以客观反映描述词的多种属性的关系。
一对多比较型	与其他……相比，……方法+[在……方面]……	与其他方法相比，MCLP更加简洁实用，它允许用户输入不同的参数对模型进行调整，具有更好的灵活性。 与其他语义相关性度量方法相比，本文方法没有涉及繁重的文本处理工作，计算量小且效果显著，尤其适合于实时性要求高的信息检索系统。

一对多比较型是指一种方法与多种方法或一类方法相比较，描述其优点或缺点，与某类方法比较时，往往会伴有“传统”“经典”“以往”“既有”“先前”“普通”“一般”等修饰词，有时也会出现几种具体的方法用于比较，会有明显的并列连词或表示并列的标点符号，如“与……方法及……方法相比较，……方法更……”。一对多的比较，往往描述优点的情况多一些。描述缺点的情况一般不使用一对多比较。在一对多比较型句子论述中，有时会出现比较的对象，有时不会出现明确的比较对象，如“与其他方法相比，MCLP更加简洁实用”，这种情况就无从抽取比较的对象了。

3.4　方法的流程描述

方法的流程描述包括方法的使用过程、方法的使用步骤、方法的使用条件等内容。方法的过程描述规则类型较多，有些描述带有明显的过程特征词，如“首先”“然后”等词。例如，“Yeung等设计了一个算法，先通过聚类方法，使对象和标签归属到不同的主题，然后考察用户的标签集，确定用户在各个兴趣主题下的标签向量。”“本文利用谱特征排列的直推式迁移分类方法对客户流失进行预测，具体流程包括：首先设计了数据维数合并以及数据属性统一方法对不同领域的数据表现形式进行统一；然后利用谱特征排列方法建立不同领域数据之间的映射关系，实现异质领域特征分布的近似统一；最后利用TSVM模型对客户数据分类从而实现对忠诚客户和流失客户的识别。”

有些方法的流程描述是针对方法的使用前提或条件，例如“该方法的前提假设是每个作者

对论文的贡献都是相同的,即有相同的贡献因子”。有些描述没有规则可循,例如“个性化排序方法,使用用户搜索历史信息训练用户兴趣模型,采用协同推荐算法获取具有共同兴趣的邻居用户,根据邻居用户对文档的推荐程度和文档与用户兴趣模型的相关程度来排序搜索结果”。这种没有明显逻辑信号词的情况,判断句子为方法的流程描述则比较困难。

整体上讲,方法的流程描述难以用单个句子来表述,往往都是由句群或段落构成。因此,构建句子级方法的流程描述规则是比较困难的。

3.5 方法的功能描述

方法的功能描述是指描述方法能解决哪种或哪类问题,对方法的应用范围或适用领域进行界定。对方法的功能描述包括以下规则:“借助|通过|使用|采用|利用|用……方法+来检验|对……进行……”,或者“……方法+能|可以+……”。其规则与举例如表 7 所示。

表 7 方法的功能描述规则及举例

规则	举例
借助|通过|使用|采用|利用|用……方法+来检验|对……进行……	使用 ADF 检验法来检验时间序列的平稳性;使用方差分析法来检验时间序列的周期性并提取周期值;使用小波估计法来检验时间序列的自相似性并计算表征自相似程度的 Hurst 参数。 使用机器学习方法对已知数据进行训练得到相关度判别规则。
……方法+能|可以+……	本方法能较为准确地揭示用户的兴趣,产生的推荐资源与用户兴趣匹配程度较高。 利用此方法可以在 Web 文本分类过程中充分挖掘 Web 文本语义信息,提高网络主题舆情分析的准确度。
……方法+可用于|适用于|被应用到+……	元胞自动机(Cellular Automata,CA)是一个典型的复杂系统研究方法,是一种局部动力学模型,适用于在空间复杂系统中的时空动态模拟研究。 目前迁移学习方法被应用到一些现实问题,如 WiFi 定位、情感分类和垃圾邮件过滤等。
……的目标是……	UIMA 的目标是提供在企业级的环境中处理各类非结构化的信息资源的通用解决方法和支撑技术。
……方法+在……领域得到应用	关联规则挖掘主要研究关系数据中关联规则的发现,根据设定的置信度和支持度阈值,挖掘出不同属性数据之间的依赖关系,这是目前处理海量数据的一种常用方法,在多个领域得到广泛应用,如网络故障检测、气象信息分析和医学病理分析等。
实验证明了|实验结果表明……方法……	实验证明了方程组度量知识创造和确定知识创造的生命周期的数量方法的实用价值。

4 结论与讨论

本文将方法知识元总结为方法定义知识元、方法关系知识元、方法特点知识元、方法流程知识元和方法功能知识元五种类型。通过半自动的方法,初步构建了五种方法知识元的描述规则,并给出一些详细示例。这些知识元的类型剖析以及描述规则,有助于后续的方法知识元抽取的技术实现。

研究中发现,不同类型的知识元有着不同的描述方式,句子复杂度与描述的类型也有较大差异。方法定义知识元相对简单,往往都是一个句子,而且句子规则相对简单。方法特点

知识元描述优缺点较多，中性描述偏少，对于优缺点的论述，多使用比较句与其他方法进行一对一比较或一对多比较，句法规则相对比较复杂。方法之间的关系包括空间静态关系与时间上的动态关系，静态关系的类型不多，但规则的描述较为复杂；动态关系的类型较多，但规则的描述并不复杂，而且，有些关系（如替代关系）难以在某个句子或某篇文章中显性地表示出来，规则的构建也比较困难。方法的流程虽然有些连接词可以辅助判定，但大部分难以用一两个句子来描述，句群或段落的描述更加适合，所以本文的这种方法也难以构建出适用的规则。方法的功能描述可以细分为领域应用、问题解决等多个方面的描述，一般可以用句子完成论述，规则的构建并不难。

当然，本文的研究还是初步的，存在一些问题或不足。一方面，目前没有专门针对方法抽取的测试集，抽取实验结果难以测评，每种方法知识元的规则数量多少为宜，也缺乏相应的评估标准；另一方面，由于语言的复杂性与不同学科之间论文风格的差异性，规则存在抽象度不高以及覆盖度不足等问题。作者将在后续的研究中增加规则抽取的原始语料规模，并构建一部分方法知识元的测试集，以验证利用方法知识元规则进行知识抽取的效果。

参考文献

[1] 温有奎，焦玉英.基于知识元的知识发现[M].西安：西安电子科技大学出版社，2011.（Wen Youkui，Jiao Yuying. Knowledge discovery based on knowledge element[M]. Xi'an：Xidian University Press，2011.）

[2] 高继平，丁堃，潘云涛，等. 知识元研究述评[J]. 情报理论与实践，2015(7)：134-138，133.（Gao Jiping，Ding Kun，Pan Yuntao，et al. Review on study to knowledge element [J]. Information Studies：Theory & Application，2015(7)：134-138，133.）

[3] 姜永常.基于知识元语义链接的知识网络构建[J].情报理论与实践，2011(5)：50-53，45.（Jiang Yongchang. Constructing knowledge network based on semantic link of knowledge element [J].Information Studies：Theory & Application，2011(5)：50-53，45.）

[4] 文庭孝，侯经川，龚蛟腾，等.中文文本知识元的构建及其现实意义[J].中国图书馆学报，2007(6)：91-95.（Wen Tingxiao，Hou Jingchuan，Gong Jiaoteng，et al. Construction and values for knowledge element of Chinese text [J]. Journal of Library Science in China，2007(6)：91-95.）

[5] 王宇，刘森.一种基于知识元的期刊文献知识仓库构建[J].情报理论与实践，2013(8)：91-94.（Wang Yu，Liu Miao. Constructing knowledge warehouse of journal literatures based on knowledge element[J].Information Studies：Theory & Application，2013(8)：91-94.）

[6] 姜永常，杨宏岩，张丽波.基于知识元的知识组织及其系统服务功能研究[J].情报理论与实践，2007(1)：37-40.（Jiang Yongchang，Yang Hongyan，Zhang Libo. Knowledge organization and services based on knowledge element [J].Information Studies：Theory & Application，2007(1)：37-40.）

[7] 仲秋雁，郭艳敏，王宁，等.基于知识元的非常规突发事件情景模型研究[J].情报科学，2012(1)：115-120.（Zhong Qiuyan，Guo Yanmin，Wang Ning，et al. Research on unconventional emergency scenario model based on knowledge element [J]. Information Science，2012(1)：115-120.）

[8] 温有奎，温浩，徐端颐，等.基于创新点的知识元挖掘[J].情报学报，2005，24(6)：663-668.（Wen Youkui，Wen Hao，Xu Duanyi，et al. Knowledge element mining in knowledge management[J]. Journal of the China Society for Scientific and Technical，2005，24(6)：663-668.）

[9] 冷伏海,白如江,祝清松. 面向科技文献的混合语义信息抽取方法研究[J]. 图书情报工作,2013(11):112-119.(Leng Fuhai, Bai Rujiang, Zhu Qingsong. A hybrid semantic information extraction method for scientific research papers [J] Library and Information Service,2013(11):112-119.)

[10] 周宁,余肖生,刘玮,等.基于XML平台的知识元表示与抽取研究[J].中国图书馆学报,2006(3):41-45.(Zhou Ning, Yu Xiaosheng, Liu Wei, et al. Study on representation and extraction for knowledge element based on XML[J]. Journal of Library Science in China, 2006(3):41-45.)

[11] 朱丽萍,李洪奇,杨中国,等. 一种面向科技文献引言的信息抽取方法[J]. 山东大学学报(理学版),2015(7):23-30,37.(Zhu Liping, Li Hongqi, Yang Zhongguo, et al. An information extraction method for scientific literature introduction[J].Journal of Shandong University(Natural Science) ,2015(7):23-30,37.)

[12] Ultsch A. Knowledge extraction from self organizing neural networks[EB/OL]. [2006-09-14]. http://www.mathematik.uni-marburg.de/~databionics/de//downloads/papers/ultsch93knowledge.pdf.

[13] 谢孟军,黄国兴,蔡健.基于Rough Set的规则自动抽取设计方案[J].计算机工程,2002(3):167-168,213.(Xie Mengjun, Huang Guoxing, Cai Jian. Approach of rule automatic extraction based on rough set [J]. Computer Engineering, 2002(3):167-168,213.)

[14] 孙晨,周志华,陈兆乾.神经网络规则抽取研究[J].计算机应用研究,2000(2):34-37.(Sun Chen, Zhou Zhihua, Chen Zhaoqian. Extracting rules from neural network [J]. Application Research of Computers,2000(2):34-37.)

[15] 侯广坤,张劲峰.基于决策树的神经网络规则抽取方法[J].中山大学学报(自然科学版),2000(4):27-30.(Hou Guangkun, Zhang Jinfeng. A rule extraction method based on decision tree[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni, 2000(4):27-30.)

[16] 高阳,陆鑫,李宁,等.一种自适应概率规划规则抽取算法[J].南京大学学报(自然科学版),2003(2):145-152.(Gao Yang, Lu Xing, Li Ning, et al. An adaptive rule extracting algorithm in probabilistic plan [J]. Journal of Nanjing University (Natural Sciences), 2003(2):145-152.)

[17] 化柏林. 针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术, 2013, 29(6): 68-75.(Hua Bolin. Extracting information method term from Chinese academic literature[J]. New Technology of Library and Information Service, 2013, 29(6): 68-75.)

化柏林 北京大学信息管理系助理教授。北京 100871。

(收稿日期:2015-12-22)