广西科学院学报
Journal of Guangxi Academy of Sciences

2018,34(1):46～50
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-03-20 DOI:10.13657/j.cnki.gxkxyxb.20180320.004
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180320.1021.008.html

# 知识图谱概念获取研究进展*

# A Survey of Knowledge Graph Concept Extraction Methods

边慧珍,哈 斯**
BIAN Huizhen, HA Si

(内蒙古师范大学计算机信息与工程学院,内蒙古呼和浩特 010022)
(Computer & Information Engineering College, Inner Mongolia Normal University, Hohhot, Inner Mongolia, 010022, China)

**摘要**:随着 Web 技术的不断更新与发展,知识图谱以其强大的语义处理能力与开放互联能力吸引了各行各业的关注。各行各业都在纷纷构建所属领域的知识图谱,如何从不同数据源抽取构建知识图谱所需概念,成为知识图谱构建的关键技术,概念抽取得越完整,所构建的知识图谱越全面,利用价值越高。本文对不同数据源抽取知识图谱概念进行阐述说明,以期引导学者选择合理的方法进行学术分析,提升知识图谱应用水平。

**关键词**:领域知识图谱 概念抽取 数据源
**中图分类号**:TP391.1 **文献识别码**:A **文章编号**:1002-7378(2018)01-0046-05

**Abstract**: With the continuous updating and development of Web technology, the knowledge graph has been favored by various fields with its powerful semantic processing ability and open interconnection ability. The various walks of life are building the knowledge graphs in their own fields. How to extract required knowledge concepts from different data sources to build the knowledge graph becomes a key technology. The more complete the concept extraction, the more comprehensive the knowledge map, the higher the value of the use. This paper expounds the concept of the knowledge graph extracted from different data sources in order to guide scholars to choose a reasonable method for academic analysis and enhance the application level of the knowledge graph.

**Key words**: concept extraction, domain knowledge graph, data source

## 0 引言

知识图谱由 Google 率先提出,其开发的目的主要是用于对真实世界中存在的各种实体和概念,以及这些实体、概念之间的关系进行研究。国外知识图谱的相关研究工作已经开展很多年,取得了丰硕的成果,也产出了大量优秀的知识图谱,其中包括 DBpedia、YAGO、FreeBase、Google Knowledge Graph 等。这些知识图谱已经在知识搜索和自然语言处理等领域发挥了重大的作用,Google 知识图谱的构建更是基于知识图谱开创了新的智能搜索模式。如今国内也有不少团队在进行这方面的研究工作,例如 XLore、Zhishi.me 等都是优秀的中文知识图谱项目。知识图谱与本体最大的不同在于,前者

**收稿日期**:2017-10-20
**修回日期**:2017-11-20

**作者简介**:边慧珍(1991－),女,硕士研究生,主要从事语义网与蒙古文信息处理研究。

*国家自然基金项目"蒙古语词汇语义网研究"(61363035)资助。

**通信作者:哈 斯(1976－),男,博士研究生,教授,主要从事语义网与蒙古文信息处理研究,E-mail: hasi76@126.com。

在实体层面上对后者进行了丰富和扩充;本体在构建过程中特别强调的是概念以及概念之间的语义关联关系,为知识图谱的数据模式奠定了基础;而知识图谱则是在本体构建的基础上丰富了概念的信息。知识图谱广泛应用于数据挖掘、智能问答、语义搜索、知识工程等领域。因此如何在已有数据源的基础上,最大可能且准确地抽取构建领域知识图谱的概念,是知识图谱构建需最先解决、也是最重要的一步。

本文详细描述了分别从非结构化数据、结构化数据以及半结构化数据中,抽取构建领域知识图谱所需概念的方法,利用这些数据不同的结构特征,尽可能保证所构建的知识图谱的覆盖面,提高知识图谱的质量。

## 1 非结构化数据的概念抽取

非结构化数据一般指没有固定结构的数据,本文指的非结构化数据是指纯的文本,文本信息广泛存在于互联网中,是知识图谱概念抽取十分重要的来源之一。由于其结构的不特定性,机器无法解读其包含的语义信息,因此,从非结构化的文本中获取知识,一般需要通过自然语言处理(Natural Language Processing,NLP)技术进行预处理操作,包括分词、词性标注、命名实体识别和句法分析等步骤,然后依靠统计分析、机器学习等方法来获取知识。

Text-to-Onto 是一个相对较早的本体学习系统,Text-to-Onto 首先采用加权词频统计方法来获取概念。Shamsfard 和 Barforoush[1] 提出了一种与领域无关的概念抽取方法,该方法基于一个简单的本体内核(包含少部分概念和关系),然后使用文本理解相关技术自动抽取本体中的概念。Lee 等[2] 则提出一种基于片元(Episode-based)的本体构建方法从无结构的文本文件中学习领域本体,其中概念的聚类与分类主要使用了基于模糊数(Fuzzy numbers)的概念相似度计算。Navigli 和 Velardi[3] 提出了一个基于文本的领域本体学习工具 OntoLearn,该工具能够从网站的文档中自动获取概念以及概念之间的关系,其最主要的特点是使用了语义解释(Semantic interpretation),即把一个复杂的概念(Concept)与一个复杂的属于(Term)相关联;通过使用 WordNet 作为通用本体,把所抽取的术语关联到 WordNet 中的概念进行解释,从而判断术语之间的分类学关系,进而确定哪些术语能被划分为概念。

本节介绍 3 种通过对文本进行处理来自动获取领域概念,以及构建知识图谱的方法。

### 1.1 基于语言学的抽取方法

基于语言学的方法就是利用预先建立的规则进行概念的抽取,该方法要求必须对领域语料集中的词法结构有精准的分析,概括出一定的词法规则,依据总结的规则建立匹配模板;然后利用这些模板对领域文本进行匹配操作以抽取概念。这种方法依赖具体语言建立相应的语法规则,必须对该领域语料集的语言环境进行学习与处理。

由于基于语言学的方法是在大量文档分析中得到语言学规则,根据获取的规则匹配文本比较简易,计算量小,概念提取的准确度较高,对于符合匹配规则但出现频率较低的词也能够抽取出来。用语言学的方法进行领域概念的抽取过程并不依赖用以建立规则的基础文档。

但这种基于语言学的方法在规则的构建上需要大量资源;抽取的结果在很大程度上取决于所构建的规则;规则的维护比较困难;对语言的依赖性很强,建立的规则不能通用于任何领域,对于不同的领域需构建不同的语法规则,可移植性较差。

### 1.2 基于统计的抽取方法

基于统计的抽取方法[4] 依据领域概念与普通词汇在文档集中具有的不同统计特征识别目标领域的概念。基于统计的抽取方法比语言学的方法更易于实现,且不局限于某一特定的领域,是目前主流的概念抽取方法。一般都把在主题领域文档内出现的相对高频率的词代表领域特征的词和短语。

利用领域相关度和领域一致度相结合的概念抽取方法,用非目标领域的文档集过滤了大量与所构建领域无关的术语,是一种典型的基于统计的概念抽取方法。其中领域相关度反映术语与指定领域的密切相关程度。领域一致度反应术语在主题领域文档集合中的分布情况,分布越均匀,则该术语代表概念的概率就越大。

在一个指定的领域文本集合中术语出现的频率也代表概念在相关指定领域出现的频率。可以采用与指定领域无关的文本集合做支撑,通过利用领域相关度与领域一致度相结合的方法更准确的提取领域文本中的概念。

基于统计的抽取方法不需要根据具体句法语义信息,不局限于某一具体的领域,不受限于语言类型和句型结构,易于扩充。但是,这种方法需要大量的领域文档做支撑,计算量大;低频术语抽取效果比较差,概念抽取的结果准确度一般达不到预期值。

### 1.3 混合方法的抽取方法

在实际的知识图谱构建过程中经常将统计方法与语言学的方法相结合起来进行概念抽取，采用规则匹配得到候选概念后用统计方法进行过滤，或者先用统计的方法得到候选概念，然后用语言学的方法建立规则来筛选领域概念，使抽取出的概念具有较高的准确度与高覆盖率。

## 2 结构化数据的概念抽取

结构化数据是知识图谱的一个重要数据来源，通过仔细研究发现，目前信息系统中大量数据都是以结构化的数据源形式存储在关系数据库。因此可以从数据库中自动提取领域知识图谱所需概念。目前知识图谱广泛采用的结构化数据有 DBPedia、Yago 等同用语义数据集，以及 Music Brainz、Drug Bank 等特定领域知识库。在使用方面，由于它们本身具备结构，通过简单的配置解析后，即可轻易地把关系模型中的模式信息和数据信息提取出来。

### 2.1 采用 RDF 模式获取结构化数据中的知识

RDF 是一种用以描述资源的框架，可以将结构数据形式化地表示，它一般用来描述网络资源，例如某个 web 页面内容、作者等。通过 RDF 对知识进行结构化组织，然后采用图形化的方式展示出来。曾锦麟采用 RDF 对语义知识进行表示，并将其应用到网上求职招聘系统中，构建网上招聘系统的本体模型[5]。

D2R 是一种基于 XML[6] 的语言，它可以将结构化数据转化为知识的 RDF 描述，从而将这些数据导入到知识图谱中[7]。

### 2.2 人工构建的本体或辞典概念抽取方法

人工构建的本体或辞典是机器可读的，一般被视为结构化数据。具有良好的实用性和可靠性。WordNet[8]、Cyc[9] 与 OpenCy 是之前本体构建用的最多的通用本体，这些通用本体有时候被用于直接的概念抽取数据源，有时也与其它数据源结合抽取概念，有时也作为启动过程的初始数据，还有些情形下用于做学习结果评估。对于一些缺乏大型知识库但有自己语义词典的少数民族语言，如蒙古语[10]，其有对应的蒙古文 WordNet，若要构建蒙古文领域本体[11]、知识图谱，蒙古文 WordNet 是其主要的数据来源。

在此主要以 WordNet 为例来说明基于开放本体或辞典来抽取构建知识图谱所需概念的方法。本文提到的 WordNet 是由东南大学提供的数据表格。

WordNet 是由 Princeton 大学的语言学家、心理学家和计算机工程师联合设计的一种基于认知语言学的英语词典[12]。它不是光把单词以字母顺序排列，而且按照单词的意义组成一个“单词的网络”。与普通的字典相比，WordNet 的最大特点是包含一定的语义。WordNet 中同义词集合的数量在 11.7 万左右，不同的同义词集合之间还存在语义关系，最主要的关系为上下位关系。

WordNet 作为一个通用本体，几乎涵盖了所有领域的概念与概念间的关联关系，从 WordNet 中抽取所需概念和关系不仅使资源得到充分利用，而且与从领域文档或互联网中获取概念和关系的技术相比，减少了信息检索、词性标注等有关操作，节省了大量时间，提高了知识图谱的构建效率。一般认为从 WordNet 中抽取构建知识图谱概念分为以下 5 步（图 1）：

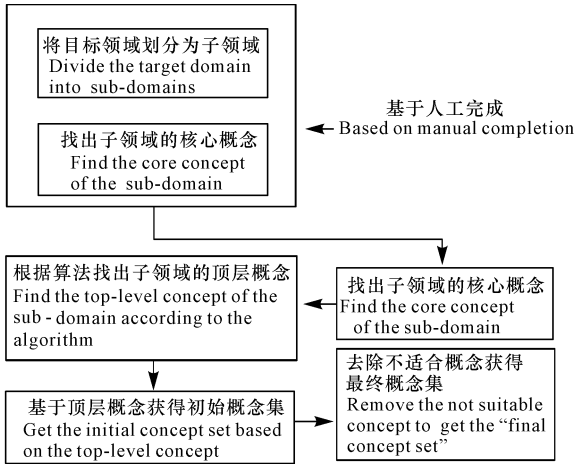


图 1 从 WordNet 中抽取概念的流程

Fig. 1 The flow chart of extracting concepts from WordNet

第 1 步：将所要构建的领域称为“目标领域”，将目标领域细分出的领域称为“子领域”。对于目标领域，根据其实际情况，结合领域书籍、维基百科以及领域相关工作者的意见，将其细分为若干个子领域（基于人工完成）。

第 2 步：找出子领域的核心概念。

第 3 步：对于每个子领域，根据核心概念，通过一定的语义相似度算法，获得每个子领域的顶层概念。

第 4 步：基于顶层概念，从 WordNet 中抽取出每个子领域所包含的概念，这样就得到了子领域的“初始概念集”。

第 5 步：去除每个子领域初始概念集中不合适的概念，来获得每个子领域所需要的概念，获得子领域的“最终概念集”。

## 3 半结构化数据的概念抽取

半结构化数据的特点是具有一定的隐含结构,但缺乏严格且固定的数据模式,互联网中大量的以HTML格式存在的网页文档和部分以XML形式存在的文档是最具代表性的半结构化数据。对这类半结构化数据应该最大限度地利用当中的结构化信息,对于当中无结构化的数据可以使用基于文本的概念抽取方法。Sanchez和Moreno开发了一种使用搜索引擎获取与初始关键词相关的网页,然后依据概念出现的频率,从中抽取出新的概念以及概念之间的关系。有学者提出了一种从XML格式的文档中获取概念之间分类学关系,首先从文档中提取出表示文档内容的关键词,然后在这些关键词的基础上使用聚类技术,将文档按照内容的相似度分成不同的组,然后从中抽取出概念以及概念之间的关系。

下面介绍从百科中提取概念的两种方法。

### 3.1 将类别标签作为概念的候选

一般认为百科中位于分类体系中的类别标签都直接视为概念,因为百科的分类系统基本都经过百科管理员以及高级编辑人员的检验,具有非常高的可靠性。

其次,其它类别标签也是概念的候选,但是不能直接选取为概念,因为在百科中存在不合理的类别,包括:(1)既没有子类别也没有论述它的文章的类别(空类别标签)。对于这种类别标签,直接把它们看成实体(在知识图谱构建的最底层);(2)没有子类别的标签,但包含以自己为标题的文章页面。同样,这些标签直接被认为是实体,其所属父类则设置为实体的概念。

概念还可抽取自上下位关系,当把最终得到的上下位关系组织成一个分类系统时,不是位于最底层(叶子节点)的均视为概念。

### 3.2 将标题作为主概念,相关概念作为概念库的候选概念

网络百科是构建知识图谱常用的数据来源[13]。

网络百科是大规模的开放式百科全书,具有质量高、覆盖范围广,实时更新和半结构化等特点,是用以构建知识图谱的优质语料来源。图2显示了将标题作为主概念,相关概念作为概念库的候选概念的概念抽取流程。

下载最新的网络百科XML语料集合,如http://download.wikimedia.org/zhwiki/latest,所有百科概念对应的解释文档包含在<doc></doc>标签中。<doc></doc>标签中的文档片段包含了概念的解释摘要和链接信息等有关信息。

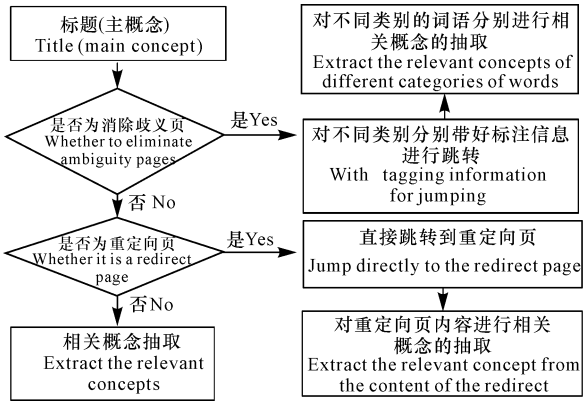


图2 网络百科概念抽取判断流程

Fig. 2 The flow chart of Encyclopedia concept extraction

概念抽取前,需把标题(主概念)在维基百科页面中进行搜索并进行逻辑判断,以提高知识图谱的质量[14]。研究表明维基百科是通过包含在页面的链接来呈现相关概念的,将主概念页面正文中出现过的链接概念作为相关概念的来源。利用语义相似度计算方法分别计算标题(主概念)与相关概念的相似度,若候选概念在语料中不存在解释片段即没有对应的摘要,认为其相似度为零。

对(标题)主概念进行逻辑判断以后,通过超链接对相关概念进行抽取,在预处理模块中可以得到一系列的相关候选概念,但是在抽取出的候选概念中仍然存在着一些概念与主题相关性不大,但它又不属于冗余信息,对于这种弱相关概念,人类大脑可以通过思维理解能力进行判断,机器却无法自动做出强弱性判断,因此本文提出了两种匹配模式对候选相关概念进行进一步的筛选。

(1)基于语言学特征(词性和词频)进行概念筛选

对弱相关概念根据相关概念的词性以及出现在正文中的频率对其进行综合加权打分,以此来确定其是否可加入到领域概念库。我们也可以通过构建名词词组解析方法来判断候选概念是否可作为概念。基于语言学特征的方法能够较好的提高精度,但是面对海量的数据时在进行计算处理操作时其复杂度较高,可移植性较差。

(2)基于已有知识库的概念名称筛选

对于弱相关概念我们可以通过设计算法计算其与主概念的相关度以及相似度,如果相关度以及相似度达到设定的阈值就将其加入到领域概念库中。

## 4 结束语

本文对中文知识图谱的概念抽取进行了探讨，总结以及提出了一些概念抽取的方法，论述了从不同数据源间抽取概念的方法。在当前互联网大数据的背景下，如何有效的组织和利用结构化数据、半结构化数据和非结构化数据等各种类型的数据，使这些数据更好的为知识图谱构建服务，成为大数据时代的新挑战。

知识图谱在语义检索、数据挖掘、人工智能、知识组织和智能问答等领域的应用非常广泛[15]。各行各业与知识图谱的深度结合，使数据信息变得更有价值，为相关科学领域开启新的发展方向。

知识图谱的构建是一个工作量巨大且比较困难的工作，虽然目前对知识图谱的研究工作有了很多有意义的尝试，但总的来说还不够完善和深入，后续还有很多研究工作需要完善，比如知识对齐、知识融合、知识图谱的绘制、知识推理以及质量评估。

**参考文献：**


[1] SHAMSFARD M, BARFOROUSH A A. Lerning ontologies from natural language texts[J]. J Human-Computer Studies, 2004, 60(1): 17-63.

[2] LEE S C, KAO Y F, KUO Y H, et al. Automated ontology construction for unstructured text documents[J]. Data & Knowledge Engineering, 2007, 60: 547-566.

[3] NAVIGLI R, VELARDI P. Learning domain ontologies from document warehouses and dedicated websites[J]. Computational Linguistics, 2004, 30: 151-179.

[4] 杜波，田怀凤，王立，等. 基于多策略的专业领域术语抽取器的设计[J]. 计算机工程，2005，31(14)：159-160.
DU B, TIAN H F, WANG L, et al. Design of domain-specific term extractor based on multi-strategy[J]. Computer Engineering, 2005, 31(14): 159-160.

[5] 杨思洛，韩瑞珍. 国外知识图谱绘制的方法与工具分析[J]. 图书情报知识，2012(6)：101-109.
YANG S L, HAN R Z. Analysis of foreign methods and tools of mapping knowledge domain[J]. Documentation, Information & Knowledge, 2012(6): 101-109.

[6] 袁旭萍. 基于深度学习的商业领域知识图谱构建[D]. 上海：华东师范大学，2015.
YUAN X P. Construction of business knowledge graph based on deep learning[D]. Shanghai: East China Normal University, 2015.

[7] HAO J, YAN Y, GONG L, et al. Knowledge mapbased method for domain knowledge browsing[J]. Decision Support Systems, 2014, 61: 106-114.

[8] 姚天顺，张俐，高竹. WordNet综述[J]. 语言文字应用，2001(1)：27-32.
YAO T S, ZHANG L, GAO Z. Introduction of WordNet[J]. Applied Linguistics, 2001(1): 27-32.

[9] ZHOU Y H, FENG J Y, YOU L S, et al. Matrine and CYC116 synergistically inhibit growth and induce apoptosis in multiple myeloma cells[J]. Chinese Journal of Integrative Medicine, 2015, 21(8): 635-639.

[10] 那顺乌日图. 蒙古语语言知识库的建立与应用[J]. 中文信息学报，2011，25(6)：162-165.
NASUN-URT. Construction and application of Mongolian language knowledge bank[J]. Journal of Chinese Information Processing, 2011, 25(6): 162-165.

[11] 杨家奇. 基于WordNet英汉蒙三语领域本体的构建方法研究[D]. 呼和浩特：内蒙古师范大学，2016.
YANG J Q. Research on the construction method of the domain ontology based on WordNet in English and Chinese Mongolian three languages[D]. Hohhot: Inner Mongolia Normal University, 2016.

[12] 李便霞. WordNet应用问题研究[J]. 科技致富向导，2013(23)：280.
LI B X. WordNet application problem research[J]. Guide of Sci-tech Magazine, 2013(23): 280.

[13] 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[J]. 计算机应用研究，2011，28(8)：2807-2811.
ZHANG H S, MA D M, DENG Z L. Semantic knowledge bases construction based on Wikipedia[J]. Application Research of Computers, 2011, 28(8): 2807-2811.

[14] 苏小康. 基于维基百科构建语义知识库及其在文本分类领域的应用研究[D]. 武汉：华中师范大学，2010.
SU X K. Research on building Wikipedia semantic knowledge base and its application in text classification[D]. Wuhan: Central China Normal University, 2010.

[15] 知识图谱技术原理介绍[EB/OL]. [2017-10-21]. https://wenku.baidu.com/view/b3858227c5da50e2534d7f08.html.
Introduction to knowledge technology principles[EB/OL]. [2017-10-21]. https://wenku.baidu.com/view/b3858227c5da50e2534d7f08.html.


（责任编辑：米慧芝）