《智能信息处理》课程作业

基于本体的语义挖掘模型构建

郭学超

作业	分数[20]
得分	

2020 年 12 月 06

第 25 卷 第 2 期 计 算 机 学 报 Vol. 25 No. 2
2020 年 12 月 CHINESE JOURNAL OF COMPUTERS Dec. 2020

基于本体的语义挖掘模型构建

郭学超

（大连海事大学 计算机技术 辽宁省大连市 中国 116026）

摘 要： 建立高效的 Web 挖掘模型是解决信息化时代数据量大、增值速度快、复杂、异构问题，并提升挖掘的工作效率以及所获取知识规则的有效方法。文章借鉴领域本体概念，通过将本体映射技术与数据挖掘技术相结合，在 Web 环境下构建了基于本体映射的 Web 语义挖掘模型，并对最终获取的知识模式在语义方面进行修正。

关键词： 本体映射；语义 Web；语义挖掘

中图法分类号： TP311.20 **DOI 号：** 10.3969/j.issn.1001-3695.2014.01.030

Building semantic mining model based on Ontology

Abstract： Building an efficient Web mining model is an effective way to solve the problems of large amount of data, fast value-added, complex and heterogeneous in the information age, and to improve the efficiency of mining and the knowledge rules obtained. In this paper, the concept of domain ontology is used for reference. By combining ontology mapping technology with data mining technology, a web semantic mining model based on ontology mapping is constructed in the web environment, and the semantic aspects of the final knowledge pattern are modified.

Key words： Ontology Mapping; Semantic Web; Semantic Mining

0 引言

“本体论”最早是哲学中的基本概念，它是研 究“ 是” 之所以为“是”的理论,可以说是哲学中 的哲学,甚至可以认为西方哲学自身的发展就是 一个“本体论”的产生、发展、怀疑和批判的过 程。近年来，本体论的方法在知识工程领域得到 了越来越广泛的应用，在很多有名的知识系统中， 如美国 D.Lenat 教授领导研制的大型常识知识库 系统 Cyc,Pricnteno 大学 Berkeley 分校研制的 语言知识库 WrodNet 等,本体论都有一定的应用。 一方面,本体论研究深层次上的指示,把知识工程研究中的知识向更深更本质的方向上推进,另一 方面,本体论的研究独立于任何语言,因此本体论 将会为不同系统之间知识的共享和互操作提供手段。

早在 1998 年,Gruber 就已经给出了本体的一个流行定义,即“本体是领域概念化对象的明确 表示和描述”。Guarino 把概念化对象 C 定义为:C<D,W,R>,其中 D 是一个领域，W 是该领域中 相关的事务状态集合，R 是领域空间<D,W>概念关 系的集合。因此，从概念化对象的定义来看，本体把现实世界中的某个领域抽象成一组概念(如 实体、属性、进程等)及概念间的关系。某个领域 的本体不仅提供了关于该领域的一个公认的概念集，同时也表达了各概念间所具有的各种语义联 系。 随着数据挖掘技术在商业领域中得到越来越 广泛的应用，对数据挖掘算法以及方法的研究也 日新月异，有关数据挖掘过程各阶段的新思想、 新算法、新技术层出不穷。一个简单，但典型的 数据挖掘过

程可能包括数据预处理阶段，数据挖掘算法的应用阶段，以及对挖掘结果可视化处理 阶段。由于数据挖掘是包含多个阶段的知识发现 过程，而在每个阶段都会有多个算法或方法供数据挖掘工作者选择，但仅有一些算法和方法组合 是有效的。因此，即使是数据挖掘领域的专家， 在一个具体的挖掘任务进行到某一个阶段时，也 难免会产生困惑：该阶段可用的技术有哪些？这 些现成的技术是否合适？若不合适，采用的新技 术以后能否被其他研究者使用？ 产生的结果是不是用户最需要的？ 基于上述原因，在本文中我们将本体的概念 引入到数据挖掘方法中，不同于其他基于本体论 的数据挖掘方法使用本体来表示领域知识，我们 是为已经存在的、被证明可以有效使用的数据挖 掘技术建立本体。通过数据挖掘方法本体 ，协助 不论是数据挖掘领域的新手还是专家在实施数据 挖掘过程中对众多可供选择的算法和方法进行选择[1]。

1 相关研究

1.1 语义Web的研究

语义 Web 是一种从语义层面理解词语、概念以及它们之间逻辑关系的智能网络， 使人机交互成为可能。 WWW 之父 TimBerners-Lee 在 20 世纪 90 年代末期对语义 Web 的诠释认为，它实际上相当于一种基于各种技术与知识表现的综合。20 世纪 60 年代末期， Collins、Quillian、Loftus 等人开创了语义 Web 研究的先河。Simon、Schamk、Minsky 等学者也不断地提出一些理论上的研究成果。在我国，2002 年语义 Web 技术被纳入“国家高技术研究发展计划”重点技术之列，得到了政府的大力支持和援助。随着 XML、RDF、Ontology 等语义网关键技术的成熟，其在人工智能领域的应用也更加普遍。具有代表性的人机互动工具真正意义上实现了由计算机通过“智能代理”将人类从各种繁琐的工作中解放出来。

1.2 面向 Web 的语义挖掘研究

澳大利亚 Griffith 大学开展的 WebKB 项目，借助 WordNet 设计了以本体论为主导的 Web 语义检索系统，清除句子歧义、进行词汇拓展，提升用户信息检索的准确度[2]。德国 Karlsruhe 大学设计的 On-to- broker 系统主要是使用语义恰当的标签，使整个 Web 页面有良好的结构，Web 页面有含义，方便人机理解，从而实现用户检索信息的自动推断。基于语义 Web 技术和相关理论， Plumbaum 提出了一种运用 JavaScript 引擎跟踪用户与 Web 站点会话的新思路，用来挖掘高质量的用户信息[3]。纪明奎与黄丽霞构建了一个把语义 Web 作为核心的个性化信息检索模型，能够为用户提供符合自身个性化信息需求的全方位资讯服务[4]。赵良和张云婧探究了以语义层次为基础的 Web 个性化资源推荐的方法，具体分析了确定 Web 页面重要度的途径，并详细阐述了 Web 个性化资源推荐的过程[5]。面向 Web 的语义挖掘模型是将传统的知识发现流程进一步优化的成果。经过将 Web 页面元素语义化处理使 Web 页面有良好的内容布局，页面元素有含义，便于计算机更容易理解，极大地改善了最终获取到的知识模式的质量。本文借鉴以往相关研究的结果，着重强调了本体映射技术对于发现 Web 中特定学科领域概念间相似程度的关键性，以及知识模式的语义修正和扩充对于提高知识模式质量的重要性。通过将本体技术与各种数据挖掘工具提供的算法相结合构建 Web 语义挖掘模型，解决 Web 环境下数据异构问题，实现知识的互享和重用以及质量的提升。

2 语义挖掘模型

2.1 数据挖掘技术

数据挖掘是从数据集中识别并提炼隐藏在其中的、有效的及最终可理解的模式进而形成高质量语义知识模型的关键过程。它不是自动完成的，需要依赖各种算法[6]。数

据挖掘需要有信息抽取、信息资源整合、数据预处理、数据形式转换、挖掘过程实施、知识模式形成、模式评价等过程。在 Web 挖掘实施阶段使用决策树方法，将 Web 资源有目的地分成 Web 文本数据、Web 链接数据、Web 使用数据，从中抽取一些有意义的、隐含的信息组合成网络文件，为数据预处理环节提供数据源。神经网络算法将信息转化具有适应性的处理元素 （神经元） 进行逻辑推理，是一种模拟人脑思维进行动态信息处理的抽象算法，在解决数据挖据问题上非常适用，主要用于数据的分类、预测以及知识模式的识别等过程。遗传算法作为一种通过模仿自然界中物种进化的基本规律来实现任意搜寻最优解的数学方法，它与 BP 神经网络算法正逐渐相互渗透和契合，通过建立两者之间的联系能够从数据库中抽取潜在有效的知识模式。关联规则算法首先从原始数据集中寻找全部高频项目组，之后根据已设定的支持度和置信度的最低临界值，选择合适的数理统计与多元分析工具给出的算法，探索各个高频项目组之间的关联规则。总之，在当今信息爆炸的时代，各种数据挖掘技术对于从海量的信息中发掘有效的信息资源仓库有很明显的促进作用。

2.2 领域本体的构建

领域本体是语义挖掘过程中所获取的知识模式准确度的一个参照，对实现数字内 容有效组织、语义检索和语义导航等具有重 要作用。因此，为了提高语义挖掘获取知识模式的准确度和有效性，首先要构建领域本体。构建领域本体的途径比较多，现在比较流行的是通过借鉴 Gruber 提出的本体构造规则以及斯坦福大学的NatalyaFNoy 和 DeborahLMcGuinness 提出的建议。领域本体的构建过程包括：① 确定本体的领域和范围，为了尽可能地降低本体构建的成本，要优先考虑重复使用已存在的本体；② 列举出领域中的关键术语、概念，并对领域中的类、类的层次结构以及类 的属性进行定义，这部分是对概念模型的描述，需要利用 OWL 描述语言并借助 Protégé＋OWL 插件的本体开发工具来完成；③ 创建实例；④ 检验和评价所构建的本体。

2.3 本体映射

本体映射是基于已有本体的一种本体学习技术，即对已经存在的本体进行集合、提取、删减等操作构建出一个新本体，或对原来本体进行优化。本体构建所具备的主观性和分散性特征，造成了在同一领域内保存有多数相互关联但又不完全一样本体的现象[7]。

由于这些本体间在语言层次、模型层次上存在不匹配的本体异构现象，从而造成了关联数据信息交互的障碍，引起相互关联的数据集在本体层上关联度降低甚至缺乏。本体映射的核心是寻找不同本体中元素的对应关系，从而实现不同本体之间的互相操作，形式上比较灵活，能更好地适应动态交互的、跨平台的、分布式的环境。本文借鉴领域本体概念并结合各种数据挖掘技术，构建了 Web 环境下基于本体映射的语义挖掘模型。在整个本体映射阶段：① 标准化所有目标本体，即将所有目标本体用同一形式来表示；② 解析本体的文档，从本体中提取出核心特征用于计算概念相似度；③ 开始相似性值的计算，并将计算得到的每对相似性值组合在一起形成一个多维的概念相似度矩阵；④ 依据相似度矩阵发现对应的映射规则和元素间的对应关系，一般包含进行映射的前提和对应的转化法则。

3 语义挖掘模型设计

本文构建的 Web 环境下语义挖掘模型的运行机理是先实施 Web 数据资源的挖掘，然后利用语义 Web 本体映射技术为语义修正和扩充提供指导，以便获取基于语义的高质量的知识规则 （见图 1）。具体过程是，先对 Web 数据源进行处理，通过各种数据挖掘工具给出的算法将 Web 数据源有目的地分成 Web 文本数据、Web 链接数据、Web 使用数据，然后分别从中抽取一些有价值的、潜在的信息组合成网络文件。由于网络文件中除了含有结构化数据外，还含

有大量半结构化乃至非结构化数据，所以需要对这些半结构化和非结构化数据实施预处理操作，以完成非结构化数据的删除以及数据形式由半结构化向结构化的变换。

目标数据库是对经过上述过程提取到的高质量的信息资源的集成。Web 挖掘主要是借助各种数据分析工具提供的算法，从目标数据库中获取隐藏在其中的知识模式的过程。在检验和评价知识模式阶段，首先结 合领域本体的概念体系及其领域属性，发现 知识模式中的概念簇和相关实例，计算知识 模式中每对概念的相似性值，组合成相似度 矩阵，进而产生相应的映射规则并发现不同 本体间元素的对应关系。然后将获取的知识 模式和领域本体进行对照，并参照领域本体 对知识模式进行语义修正和扩充，最终形成 语义知识模型。这样不仅强调本体映射技术对于提升 Web 数据挖掘最终获取的知识模式的质量具有重要的作用，还创新性地将本体映射技术与数据挖掘算法相结合，发现知识模式中概念间的关系，提高知识模式的质量。

3.1 数据预处理

网络文件中的数据作为形成知识模式的数据源，数据预处理的效率直接影响到最 终获取的知识模式的质量。其大致包括以下四个主要环节。

（1） 数据清理。数据清理作为数据预处理的首要任务，一般包含偏差检验及数据形式转换两个步骤其基本原理是借助相关技术，如数理统计方法、数据挖掘技术、预设模式规则法等，清除与挖掘过程无关或冗余的日志项，删除重复记录，纠正错误请求等。

在数据清理环节要注意以下几个因素：区分不同用户需要的信息；通过哪些信息有效识别用户会话；与知识模式表达及解释的数据项有哪些；如何筛选通过用户会话识别的 WebRoot 浏览记录。通常 HTML 页面中与Web 挖掘无关的日志记录、WebRood 的历史浏览日志记录以及有误的访问记录需要清洗。HTML 页面中与 Web 挖掘无关的记录主要包括一些多媒体文件、文本文件、CSS 样式表等。其中多媒体文件主要是 HTML 页面中的图像 （*.gif、*.jpeg、*.jpg）、声音 （*.mp3、*.midi、*.cd）、动画等被引用的资源。在进行数据清洗时，这些无关的记录可以通过查看 URL 的后缀来清除，如， 所有后缀名为 *.gif、*.jpeg、*.jpg、*. mp3、*.midi、*.cd、*.avi、*.swf、*.js、*.css 的文件都要被清除。WebRood 历史浏览日志记录主要通过检查请求页面的 URL 后缀来识别，清除所有后缀为 Robots.txt 的文件。 有误的访问记录中通常含有“Error”或“Failure”的状态码，服务器可以通过寻找 Web 日志中的状态码来清理有误的访问记录。

（2） 用户识别。用户识别是从浏览器历史访问记录中区别出对应的用户，建立用户与所浏览页面之间联系的过程。数据预处理阶段比较常用的用户识别方法是通过解析 Web 日志中的 IP 地址和 UserAgent 类型等信息，来区别同一 Web 站点上的用户。一般采用以下规则识别用户：① 一个 IP 地址只能唯一标识一个用户，也就是说，IP 地址不同代表的用户也不同；②如果 IP 地址一致，但是 UserAgent 信息 （如操作系统或浏览器类型） 只要有一个存在差异，就可以假定不是同一用户在访问该 Web 站点；③ 假如 IP 地址与 UserAgent 信息全部一致，则需要确定每个被请求的页面与历史访问页面间是否存在直接的链接，如果不存在，则假定同时有多个用户在访问该 Web 站点，如果存在，就默认被访问的 Web 站点上只有一个用户。

（3） 会话识别。用户会话是指用户使用特别指定的 IP 地址在一个具体的时间范围内访问一个站点的一连串活动。会话识别主要指把相同用户在一次浏览过程中的连续请求聚类形成有价值的 Web 页面序列。用户会话识别常用以下规则：① 新用户新会话同时产生；② 在某个用户会话中，若出现引用页面为空或者不存在的情况，则假设该用户又进行了新的会话；③ 若两个被请求的页面在时间上的跨度超出规定的上限 （一般为 30min），则假设新的会话又启动了。

（4） 路径补充。用户的历史浏览记录都被存储在本地缓存存储区中，Web 服务器在发送访问请求之前会检验本地缓存区中是否存在和被访问的 URL 相匹配的 URL。假如存在，那么访问请求不再被发送，直接从本地缓存中抽取目标页面提供给用户。通常， 一些点击率高的信息都会被存储在本地缓存区中。当 HTML 页面中 Meta 标志设定过期时，本地缓存就会失效，很多重要的访问记录被遗失。路径补充是用来填充遗失的页面引用，改善能够被区别出的用户会话，正确地描述用户的访问请求。假如用户目前请求页面与最后一次请求页面之间存在直接链接，则说明用户也许利用“Back”按钮进行后退来缓存网页读取；反之就认为本次用户会话没有调取本地缓存中的资源，且历史请求页面中，时间上最接近现时被请求页的页面即为现时访问请求的起源。

3.2 概念相似度计算

概念相似度作为领域本体中概念相似性的度量标准，能够表示概念间语义路径距离的远近程度。其值的获得是整个本体映射阶段非常关键的一步，影响着映射规则的产生和元素间关系的发现。目前，要想获得领域本体概念相似度矩阵，首先要算出领域本体中两两概念之间体现出的相似性的值[8]。计算之前，要判断这对概念是不是同义，若同义，就可以判定它们全部相同，相似性值记作 1。否则，得到该相似性值需要经历两个阶段：语义初始相似度阶段和非上下位关系相似度阶段。语义初始相似性值是根据每对概念之间的语义路径距离求得的，还可以认为是每对概念语义相似性的约定值，通常用 ISim（Ci，Cj） 表示概念的语义初始相似度。非上下位关系相似度可以理解为是基于语义初始相似度，经过分析每对概念的非上下位关系得到的，用 Simfss（Ci，Cj） 表示。计算非上下位相似性值之前，需要判断概念的关系类型是概念型还是 Datatype 型：Datatype 型的关系与数值型数据对应， 与概念无关；概念型的关系与概念对应，与数值型数据无关。通过对以上两种相似性值分别分配权重，算出它们的加权和即为每对概念的实际相似度 Sim （Ci，Cj） 的值。

定义 1：如果领域本体中一对概念 C1 与 C2 同义，则它们的实际相似度 Sim（C1，C2） =1。

定义 2：一对不是同义的概念 Ci 和 Cj 之间上下位关系表现出的语义初始相似度为

$$\mathrm{ISim}(C_i, C_j) = \begin{cases} \dfrac{\alpha(\mathrm{dl}(C_i) + \mathrm{dl}(C_j))}{(\mathrm{Dist}(C_i, C_j) + \alpha) \times 2 \times \mathrm{maxdl} \times \max(|\mathrm{dl}(C_i) - \mathrm{dl}(C_j)|,\ 1)} & C_i \neq C_j \\ 1 & C_i = C_j \end{cases} \quad (1)$$

其中，ISim（C1，C2） 表示概念 C1 和 C2 的语义初始相似度，dl （C1） 和 dl （C2）分别代表 C1和 C2所处的层次，Dist （C1，C2） 是概念 C1 和 C2 的语义路径距离，maxdl 是概念在本体中所处的最高层次。α 是参数，可以改变，由领域专家确定，一般≥0。此处，为了方便，将计算结果归一化，需要乘以该参数。

定义 3 ： 若 Rdt1 和 Rdt2 是一对 Datatype 型概念关系的关系名，则它们的相似度是

$$\mathrm{Sim}_{\mathrm{fss\_Rdt}}(Rdt_i, Rdt_j) = \begin{cases} 1 & Rdt_i = Rdt_j \\ 0 & Rdt_i \neq Rdt_j \end{cases} \quad (2)$$

定义 4：一对不是同义的概念 Ci 和 Cj 之间非上下位关系表现出的相似度 Simfss（Ci，Cj）为

$$Sim_{fss}(C_i, C_j) = \frac{\max_{p \in C_R(C_i, C_j)} (\sum_{(A,B) \in p} ISim(A, B)) + \max_{p \in C_R(C_i, C_j)} \sum_{(U,V) \in p} Sim_{fss\_Rdt}(U, V)}{\max((M1+M2),\ 1)} \quad (3)$$

定义 5：领域本体中，一对不是同义的概念 Ci，Cj 的实际相似度为

Sim （Ci，Cj） =β ISim（Ci，Cj） +γ Simfss（Ci，Cj） （4）

其中 β、γ 分别表示分配给两种相似度的权重（一般假设 β=γ=0.5），0<β<1，0<γ<1，β+γ=1，一般 β≥γ 。

3.3 知识模式的语义修正和扩充

知识模式的语义修正和扩充以概念相似度算法为基础，通过知识模式进行本体推

理，主动寻找知识模式中与领域本体概念不 一致的词汇，依据相应的规则删除词汇间关联度相对低的知识模式，并对词汇进行概念的全方位逻辑推理扩展。进行知识模式的语义修正和扩充能够提高特定学科领域的核心知识体系所体现的概念间的关联度，从而使数据挖掘最后提取到的知识模式更加准确、有用和全面。知识模式的语义修正和扩充的大致过程如下。

（1） 使用概念相似度计算公式求得知识模式各个词汇节点之间的相似性值 A。

（2） 依据领域本体概念体系，使用概念相似度计算公式交叉求解知识模式中具有上下位关系的词汇节点间的概念相似性值 B。如果 B≥A，就把变换过的概念模式集合保存到知识模式库中。

（3） 依据领域本体概念体系，使用概念相似度计算公式交叉求解知识模式中具有非上下位关系的词汇节点间的概念相似性值C，如果 C≥A，就把变换过的概念模式集合保存到知识模式库中。

4 参考文献

[1] 蔡皎洁.Web 环境下的语义挖掘模型研究[J]. 情 报 理 论 与 实 践,2015,38(05):121-124+111.

[2] Martin P, Eklund P. Embedding Knowledge in Web Documents: CGs versus XML-based Metadata Languages[J]. 2000.

[3] Plumbaum T , Stelter T , Korth A . Semantic Web Usage Mining: Using Semantics to Understand User Intentions[M]// User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, 2009.

[4] 纪明奎,黄丽霞.基于语义网的个性化信息检索模型研究[J].现代情报,2007(12):166-167+171.

[5] 赵良,张云婧.一种以 Web 语义挖掘的个性化信息推荐设计[J].电脑知识与技术,2011,7(08):1731-1733.

[6] 杨秀港.数据挖掘算法综述[J].科技经济导刊,2019,27(05):166.

[7] 王顺,康达周,江东宇.本体映射综述[J].计算机科学,2017,44(09):1-10.

[8] 翟社平,李兆兆,段宏宇,李婧,董迪迪.多特征融合的句子语义相似度计算方法[J]. 计 算 机 工 程 与 设计,2019,40(10):2867-2873+2884.