《智能信息处理》课程作业

基于形式概念分析的最优本体选择方法

王雪梦

作业	分数[20]
得分	

2020 年 11 月 11 日

基于形式概念分析的最优本体选择方法

王雪梦

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘 要 目前的物联网是基于关键字匹配的，在很大程度上阻碍了物联网智能化程度的发展。为此，学者们提出了“语义物联网”。简单地说，语义物联网是在物联网的基础上纳入语义协同。语义协同的两个最为重要的方面是基于本体的语义标注和基于本体的语义理解。为了使用户能够高效快速地获取自己所需要的本体，提高本体利用率和语义协同的运行效率，进而加快语义物联网的运行速度，需要为用户选择最适合他们需求的本体。因此，最优本体选择方法是当今语义物联网的研究热点之一，基于形式概念分析的最优本体选择方法是提高语义物联网运行效率的关键所在。

关键词 语义物联网； 本体； 形式概念分析
中图法分类号 TP18 **DOI 号**：1.3724/SP.J.1016.2014.01229

An Optimal Ontology Selecting Method Based on Formal Concept Analysis

Wang Xue-meng

(Information Science and Technology College, Dalian Maritime University, Dalian 116026)

Abstract The current Web of Things is based on keyword matching which is detrimental to the development regarding Web of Things. Accordingly, researchers proposed “Semantic Web of Things”. As far as Semantic Web of Things concerned, the information of things should be represented as ontology-based semantic annotation, and the information of things should be utilized as ontology-implemented semantic understanding. To require the ontologies the users need and improve the efficiency of Semantic Web of Things, most proper ontology for users should be selected. Accordingly, a method of selecting optimal ontology regarding Formal Concept Analysis is so crucial for Semantic Web of Things.

Keywords Semantic Web of Things; ontology; Formal Concept Analysis

1 引 言

作者通过串口通信技术，利用嵌入式软件 Qt 模拟手机等嵌入式设备，利用单片机模拟家用电器，做了一个物联网简易系统，通过演示该简易系统表明目前的物联网是基于关键字匹配的，也就是语法协同，在很大程度上阻碍了物联网智能化程度的进一步发展。为此，黄映辉和李冠宇等学者提出了“语义物联网”。

物联网实际上是网络上的产品信息，它依赖于产品信息。语言是要被事先约定好的，为了传输现有的信息，不论传输的是目前已经存在的或者是新创造的语言，提交的代码必须能够被接受。当前我们已经在物联网应用中有了常用的语言和代码，在使用时:只有无形的产品相关信息而不是物质的本身。

收稿日期：2020 年 11 月 1 日；最终修改稿收到日期：2020 年 12 月 13 日. 王雪梦，女，1997 年生，硕士，E-mail: 1378509927@qq.com，主要研究方向为智能信息处理.

物联网的语义就是物联网信息的相互交流。他的实质内涵是全球产品信息的共享系统。物联网更准确的可以被称为因特网上的产品信息。物联网的优势是“产品+信息+网络”，物联网的这三种属性严格的对，应一种产品的信息和传输方式。产品的信息必须写成电子标签，并通过一定的标准进行包装。物联网可以被看作一种特殊的语义网的应用形式，它试图实现基于语义网的智能处理和产品信息分享。这样的分类有助于进一步研究关于物联网的相关课题。

简单地说，语义物联网是在物联网的基础上纳入语义协同，语义协同的两个最为重要的方面是基于本体的语义标注和基于本体的语义理解。为了使用户能够高效快速地获取自己所需要的本体，加快语义物联网的运行速度，我们需要为用户选择最适合他们的本体。受到黄映辉老师课上的启发并结合形式概念分析的方法，本文提出基于形式概念分析的最优本体选择方法。

2 形式概念分析

形式概念分析一般指概念格。概念格（Concept Lattice）是一个以概念为元素的偏序集，它可以通过 Hasse 图可视化，其中每个节点是一个概念。概念格结构模型来源于形式概念分析（FCA）理论，是 FCA 中的核心数据分析工具，它本质上描述了对象（样本）与属性（特征）之间的关联。

1982 年，德国的 Wille R 提出了形式概念分析。它不仅是一种有效的数学工具，也是数据分析和规则提取的工具，同时也是知识处理的一种新方法。它能够帮助人们认识集合中各个元素之间的关系，用数学方法表达概念和概念层次[1]。

另外，概念格通过 Hasse 图生动和简洁地体现了这些概念之间的泛化和特化关系。因此，概念格被认为是进行数据分析的有力工具。从数据集中(概念格中称为形式背景)中生成概念格的过程实质上是一种概念聚类过程;然而，概念格可以用于许多机器学习的任务。目前，已经有了一些建造概念格的算法，并且概念格在信息检索、数字图书馆、软件工程和知识发现等方面得到应用。

形式概念分析的首要工作便是建立形式背景。形式背景是一个三元组 (D,A,R)，其中 D 为对象集合，A 为属性集合，R 为 D 和 A 之间的二元关系，即 $R\subseteq D\times A$。该三元组可以用二维表来表示[1]。

定义 1 一个形式背景是一个三元组（D,A,R），形式对象集合为 D，形式属性或术语的集合为 A，二元关系为 $R\subseteq D\times A$，对于 $d\in D$，$a\in A$，dRa 表示对象 d 有着属性或术语 a[1]。

定义 2 形式背景（D,A,R）中 $X\in D$，$Y\in A$，令 $X'=\{a\in A|(\forall d\in X)dRa\}$，$Y'=\{d\in D|\forall a\in Y, dRa\}$，$X'$是在 X 里所有对象的共有属性集，Y'是拥有 Y 里所有属性的对象集，如果 $X'=Y$ 并且 $Y'=X$，（X,Y）叫做一个概念，X 和 Y 分别称作概念的外延和内涵[1]。

通过定义 2 可以知道形式概念分为外延和内涵。形式概念的外延是指被表示为属于这个概念的所有的对象集合，形式概念的内涵是指被表示为所有这些对象所共有属性的集合。

定义 3 设$<K,\leq>$为偏序集，$D\in K$，a 为 K 的任一上界，若对 D 的所有上界 y 均有 $a\leq y$，则称 a 为 D 的最小上界，即上确界。同样，若 d 为 D 的任一下界，若对 D 的所有下界 z 均有 $z\leq d$，则称 d 为 D 的最大下界，即下确界[1]。

定义 4 设$<K,\leq>$为偏序集，如果 K 中任意两个元素都有最小上界和最大下界，则称$<K,\leq>$为格[1]。

定义 5 对于形式背景 $K=(O,A,B)$存在唯一的一个偏序集$<K,\leq>$与之对应，并且该偏序集的子集的上确界与下确界都存在，这个偏序集产生的格结构称为概念格[2]。

概念格主要用于认知计算、机器学习、模式识别、专家系统、决策分析、网页搜索等领域。近年来，概念格应用研究出现一些新领域，比如认知概念学习，规则提取，三支决策，等等。

在知识发现领域，概念格可以从关系数据中构造出来，然后从概念格上可以提取各种类型的知识，如蕴含规则、关联规则、分类规则等等;在软件工程领域，概念格可以从类库的规范说明上构造，从而对类库结构的可视化以及类库的重构和优化提供支持;在知识工程领域，概念格可以用于知识库的重新结构化;在信息检索方面，概念

格可以实现对信息的有机组织并过虑掉无用的信息。而且，有人指出概念格将会在生物和生命科学领域有重大应用。

3 本体的基本概念

理论上，本体是指一种“形式化的，对于共享概念体系的明确而又详细的说明”。本体提供的是一种共享词表，也就是特定领域之中那些存在着的对象类型或概念及其属性和相互关系；或者说，本体就是一种特殊类型的术语集，具有结构化的特点，且更加适合于在计算机系统之中使用；或者说，本体实际上就是对特定领域之中某套概念及其相互之间关系的形式化表达（formal representation）。本体是人们以自己兴趣领域的知识为素材，运用信息科学的本体论原理而编写出来的作品。本体一般可以用来针对该领域的属性进行推理，亦可用于定义该领域（也就是对该领域进行建模）。

Gruber 在计算机领域中第一次定义了本体，后来经过不同的学者完善和修改。目前多数学者认为本体是共享概念模型的明确的形式化规范说明。本体具有 4 层含义[3]。

(1) 概念化：通过抽象客观世界中一些现象的相关概念而得到的概念模型，即对概念系统所蕴含的语义结构，是对某一事实结构的一组非正式的约束规则。

(2) 明确性：概念机使用这些概念的约束都有明确的定义。

(3) 形式化：指本体是计算机可读的（即能被计算机处理的）。

(4) 共享性：指本体中体现的是共同认可的知识，反映的是相关领域中公认的概念，即本体针对的是社会范畴而非个体间的共识。

4 相关工作

4.1 基于本体构建方法的最优本体选择

目前，Perez 等研究者根据用户输入的查询而为用户单独构建特定的本体，比如 KACTUS 工程法、METHONTOLOGY 法、IDEF5 法、AFM 法和 SENSUS 等方法[4]。

这些方法虽然能够提供用户所需要的本体，但 是十分耗时，并不适合应用于未来的语义物联网。

4.2 基于本体融合方法的最优本体选择

Jones 等研究者通过 PROMPT、HCONE 和 OntoMerge 等本体融合工具把两个或者两个以上的本体融合在一起，从而结合各个本体的优势形成一个新的本体，进而满足用户输入的查询[5]。

这种方法虽然在很大程度上能够满足用户所输入的查询，但是融合后的本体有可能因为过大而影响语义物联网的运行效率。

4.3 基于遗传算法的最优本体选择

遗传算法是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的一种计算模型，同时也是一种通过模拟自然进化过程搜索最优解的方法[6]。

Arpirez 等研究者利用不同的选择运算、交叉运算和变异运算来求解用户所需的最优本体，这种方法虽然在很大程度上满足用户对本体的需求，但是目前有关遗传算法相关算子的选取还需要进一步的优化。

5 基于形式概念分析的最优本体选择

5.1 基于形式概念分析的最优本体选择思想

形式概念分析是一个有效的数学工具，下面利用形式概念分析中的形式背景和 Hasse 图来阐述针对用户的查询来选择相应的最优本体的思想。

表 1 本体属性的形式背景

	a_1	a_2	a_3	a_4	a_5	a_6	a_7 a_8	
O_1	×			×	×	×		
O_2	×				×	×		
O_3		×	×	×				
O_4	×					×	×	
O_5		×	×					
O_6		×				×		×
O_7	×	×	×	×				

表 1 是本体属性的形式背景表，其中 $O_1,O_2,O_3,O_4,O_5,O_6,O_7$ 表示 7 个本体，$a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8$ 表示 8 个属性，通过表 1 可以清晰地看到每个本体所具有的相应属性。由

于属性 a_3 出现必然属性 a_2 出现，属性 a_5 出现必然属性 a_1 和 a_6 都出现，再根据定义 2 和定义 5 可知属性 a_3 和 a_5 都不能形成一个独立的概念格。依据表 1，可以画出如图 1 的哈斯图。

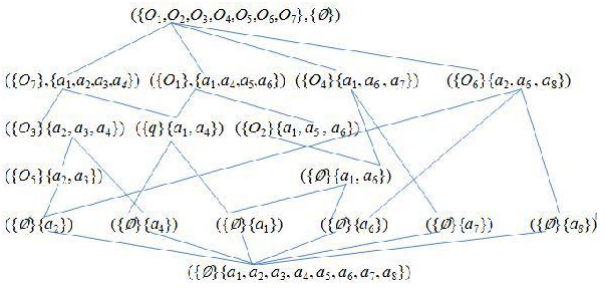

图 1 Hasse 图

图 1 中的 q 是用户输入的查询，设定 q 是一个伪本体，依据图 1 中的路径可知离伪本体 q 最近的本体是 O_1 和 O_7，路径为 1，其次是 O_2 和 O_3，路径是 2，再次是 O_4 和 O_5 ，路径为 3，最后是 O_6，路径为 4，根据离伪本体 q 的路径从小到大的顺序，有如下排序：O_1,O_7,O_2,O_3,O_4,O_5,O_6。如果两个本体离伪本体 q 的距离相同，那么本体中含有属性数量较多的排在前面，较少的排在后面。

图 1 所得到的序列是最适合查询 q 的本体到最不适合查询 q 的本体的一个排序序列，依据该序列可以选出最适合用户输入的查询 q 的本体是 O_1 和 O_7。

5.2 基于形式概念分析的最优本体选择算法

根据 5.1 所示的基于形式概念分析的最优本体选择思想，可以得到相应的最优本体选择算法。

算法 基于形式概念分析的最优本体选择算法

```
输入：一组本体
输出：针对用户输入的查询的最优本体
BEGIN
    (O,R,A)=f_constructFormalContext(O,A);//获取三元组
    conceptSet=f_constructConcept(Q); /*在三元组中获得相应的概念集*/
    L(O,A,R)=f_constructConceptLattice(conceptSet);
                        /*从概念集中建立相应的概念格*/
    q=f_getPseudoConcept(Q); //获取伪本体（查询）
    L(Oq,Tq,Rq)=L(O,A,R).addconcept(q); /*把伪本体加入到本体集中*/
  for (i=1; i<=n; i++)
  {
      r(oi)=f_getNearestNeighborPath(oi,q);
  }/*通过此循环可以把各个本体结点与伪本体结点的最短路径存储起来*/
  listRank=f_getOntologyOrder(f_getAscendOrder(r));
                        /*本体按升序排序*/
  for(i=1;i<n;i++){
      for (j=i+1;j<=n;j++){
          if(r(oi)=r(oj)&&|oi*|>|oj*|)
              listRank=f_adjustRank;
      }
  }/*如果有两个本体结点与伪本体结点距离相同，选取本体中含有术语多的那个结点*/
  return getProperOntology(listRank）; //返回最优本体
  END
```

6 方法评估

为了评估基于形式概念分析的最优本体选择方法，本文引入了对本体的人工排序方法。通过人工排序方法得到的本体序列与本文提出的方法得到的本体序列进行对比，获得相应的皮尔森相关系数，该系数表明基于形式概念分析的最优本体选择方法与人工排序方法所得到的结果的相似度大小，以此可以看出基于形式概念分析的最优本体选择方法的优劣。

表 2 基于形式概念分析的最优本体选择与人工排序比较

本体	人工排序	FCA 本体选择序列
O_1	1	1
O_2	2	5
O_3	3	2
O_4	4	6
O_5	5	7
O_6	6	3
O_7	7	8
O_8	8	4
O_9	9	10
O_{10}	10	9
皮尔森相关系数		0.879

本文选取 10 个生物学领域的本体，根据用

户输入的查询分别用人工排序方法和基于形式概念分析的最优本体选择方法进行排序，并获得相应的皮尔森相关系数，其结果如表 2 所示。

通过表 2 可知，基于形式概念分析的最优本体选择方法得到的本体序列与人工排序得到的本体序列之间的皮尔森相关系数是 0.879，由此可见基于形式概念分析的最优本体选择方法近似于人工排序的结果，其所得到的最优本体是可靠的，可以适用于以后的语义物联网。

7 小 结

基于形式概念分析的最优本体选择方法是本文提出的新方法，能够针对用户输入的查询来选择相应的最优本体，具有很强的可靠性，为以后研究语义物联网的搜索引擎、语义协同、情景计算、本体学习等内容具有极高的价值。

参考文献

[1] Engineering - Reliability Engineering; Investigators from Central University of Venezuela Have Reported New Data on Reliability Engineering (Introduction To Formal Concept Analysis and Its Applications In Reliability Engineering)[J]. Journal of Technology & Science,2020.

[2] Networks - Peer-to-Peer Networking; New Peer-to-Peer Networking Study Results Reported from School of Computing Science and Engineering (Tracing of Online Assaults In 5g Networks Using Dominance Based Rough Set and Formal Concept Analysis)[J]. Network Weekly News,2020.

[3] 彭致华,王霞.一种基于概念格的新型软件质量综合评价方法[J].电子技术与软件工程,2020(16):52-55.

[4] 李金海,魏玲,张卓,翟岩慧,张涛,智慧来,米允龙.概念格理论与方法及其研究展望[J].模式识别与人工智能,2020,33(07):619-642.

[5] 李金海,闫梦宇,徐伟华,折延宏,张文修.概念认知学习的若干问题与思考[J].西北大学学报(自然科学版),2020,50(04):501-515.