《智能信息处理》课程考试

基于本体的兴趣模型和语义相似度计算方向的研究

王 明

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 17 日

基于本体的兴趣模型和语义相似度计算方法的研究

王 明

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要 本体是语义网和语义网格研究中的一种重要方法。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。本文在基于知识本体库的基础上，提出了个性化用户模型的兴趣树构建方法，并在此基础上提出了基于 WordNeT 的语义相似度计算方法，为个性化服务提供理论指导，与其他方法相比，本文提出的方法可以为个性化用户提供更为精准的推送服务。

关键词 本体；兴趣模型；语义网；语义相似度

中图法分类号 TP391 **文献标识码** A

Research on interest model and semantic similarity computing method based on Ontology

Wang Ming

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract Ontology is an important method in the research of semantic web and semantic grid. Ontology refers to a "formal, clear and detailed description of the shared concept system". Ontology provides a shared thesaurus, that is, the object types or concepts existing in a specific field, their attributes and relationships. Based on the knowledge ontology base, this paper puts forward the interest tree construction method of personalized user model, and puts forward the semantic similarity calculation method based on WordNet to provide theoretical guidance for personalized services. Compared with other methods, the method proposed in this paper can provide more accurate push services for personalized users.

Keywords Ontology; Interest model; Semantic network;Semantic Relevancy

0 引言

随着计算机在信息化方面的应用，人们对数字信息的处理需求越来越高。信息技术开始面临知识表示、信息组织和软件复用等各种新的挑战。特别是互联网的迅速发展，使得组织、管理和维护海量信息，为用户提供有效的服务成了一个重要而紧迫的研究课题。为了适应信息化处理需求，出现了一个新的概念，本体（Ontology）作为一种能够在知识层面和语义层面描述信息系统的建模工具，引起了国内外许多研究者的关注，并被广泛应用于计算机等知识领域[1]。动态数字语义标引技术主要通过对本体资源的语义标注，充分挖掘用户感兴趣和需要的资源，为用户推送个性化的学习资源和学习计划等。主要在于建立资源领域本体，对资源进行语义标注、审核、加工等，利用本体建立实体与知识点的关联关系，知识点与资源的关联关系，最终为个性化推送服务[2]。在领域本体知识库方面，当前本体建模缺乏逻辑层知识表达的问题,一般方法是通过阐释符号与符号过程的概念,联系符号框架理论,将知识表达和关联过程分为 3 个维度,即语法、语义、语用的应用。另外一个是语义相关度计算，是信息检索、文档分类和聚类、推荐系统、机器学习等诸多领域，仍然存在一些关键技术亟待解决。

1 相关概念

本体论，最先起源于哲学领域，有的西方学家认为本体仅仅是理念，又有人认为本体是“自在之物”，而本体在计算机学科领域内，是指一

收稿日期：2021-12-20

作者简介：王明，男，1999 年生，硕士生在读，主要研究领域为自然语言处理，Email:1173037855@qq.com

种“形式化的，对于共享概念体系的明确而又详细的说明”。本体是语义网和语义网格研究中的一种重要方法[4]。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。在实践中，本体被表示成一种使用归类、部分-整体、属性-值这些关系联系起来的概念层次体系，有时也会用其他种类的关系以及被称为“公理”的规则或约束条件来扩充概念间的关系。

传统的信息检索系统大都是基于关键字匹配的检索技术，使得用户在检索时经常得到大量与查询无关的结果[5]。为了提高用户对检索系统的满意度，可以通过扩展检索系统对语义查询和动态查询的支持来现。将语义相关度计算引入信息检索技术的研究中，正是为了提高检索系统对用户查询信息的语义处理能力，从而提高检索效率，使系统更具智能性。

文本处理的核心问题。例如，语义相似度或相关度算法已经被应用于词义消歧、音频识别错误的检测、信息提取、语音自动摘要、人的姓名解析、文本相似度计算、文本分类和聚类等。

在信息检索中，语义相关度要反映的是文本或者用户查询在意义上的符合程度[4]。目前语义相关度计算的研究都是建立在对语义相似度研究的基础上的。常见的相似度计算方法 2 种，一种是根据世界知识或者分类体系计算；一种是利用大规模的语料库进行统计。本文提出了一种新的语义相似度计算方法。

2 基于本体的个性化用户模型—“兴趣树”构建方法

用户兴趣会随着主观条件和客观条件的改变而发生转变，学者们将这种转变定义为兴趣迁移，又称用户兴趣漂移。在互联网平台上，用户兴趣的转变通常可以用用户信息行为的转变来体现，即通过分析用户表征的信息(例如用户自定义标签、用户搜索历史等)来识别用户兴趣偏好。同时这些信息与用户兴趣的衰减、 增强也是相互关联的[5]

在基于本体领域资源的基础上，提出了一种基于本体的个性化用户模型——兴趣树构建的平衡方法。核心内容是：基于已经构建了信息系统的领域本体，通过领域本体中的概念关系描述用户模型，并选择用户模型中最广泛的“直接关系”和“对等关系”，从而形成用户模型，一棵树状的“兴趣树”。“兴趣树”的具体思想是在构建良好的领域本体的基础上为用户提供固定的兴趣点，然后分析用户的 URL 属性，并利用属性中包含的概念构造用户兴趣树。构建的用户模型以用户个人信息和用户兴趣树的形式呈现。用户模型的具体构建过程如图 1 所示。

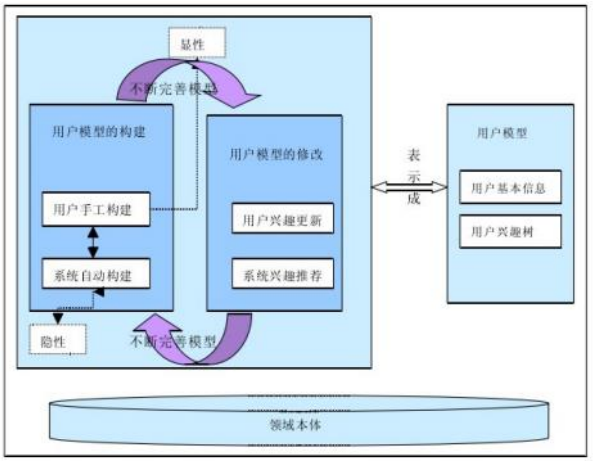

图 1 基于领域本体的用户模型构建

3 基于 WordNeT 的语义相似度计算方法

语义相似度的计算需要先确定语义信息的含义，并使用各种语义信息，包括距离、信息系数 IC(Information Coefficient)、深度、语义关系和概念特征[6]。距离是最简单、最直观的语义信息。在现有的研究中，都设置了所有的混合语义相似度。相似度的计算方法利用距离的语义信息进行计算。本文提出的计算方法也是一种距离相关计算方法。该距离分为江提出的实际物理距离和语义距离。本文提出了一种新的语义相似度计算方法：

$$Sim(c_1,c_2) = e^{-(\alpha \times L(path)+\beta \times L(IC))} \quad (1)$$

式中的参数说明如下，其中 $L(IC)$为与信息系数 IC 有关的函数，$L(path)$为最短路径距离相关的函数，和语义距离相关的函数，α 和 β 为参数，参数范围为 $\alpha > 0$，$\beta > 0$。$L(IC)$与 Jiang 定义的语义距离公式相同，即：

$$L(IC) = IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1,c_2)) \quad (2)$$

其中 $IC(c_1)$和 $IC(c_2)$为分别表示本体概念 c_1，c_2 的 IC值，$LCS(c_1,c_2)$为c_1,c_2的公共包含，$IC(LCS(c_1,c_2))$为概念c_1，c_2的最小公共包含 IC 值。公式中的所有 IC 的计算方法均使用新的 IC 计算方法。

为了确定最短路径距离 $L(path)$对语义相似度的影响，本文给出如下两种 $L(path)$的计算方法：

(1)方法 1：

$$L(Path)=\frac{Ditance(c_1+c_2)}{2*DepthMax} \quad (3)$$

(2)方法 2：

$$L(Path)=\frac{log\,(Ditance(c_1+c_2)+1)}{log\,(2*DepthMax+1)} \quad (4)$$

其中 $Distance(c_1,c_2)$表示两个概念结点 c_1,c_2的最短路径距离，$DepthMax$ 为 $WordNet$ 的最大深度。在计算 IC 的过程中，使用了诸如深度和密度等语义信息，在语义相似度的计算中使用最短路径距离和深度，因此所提出的方法属于混合语义相似度计算方法[7]。

4 个人偏好模型构建和个性化推送服务

4.1 个人偏好模型的构建

根据语义本体和相似度计算方法，个人偏好模型知识来源如图 2 所示，分为 5 个步骤：

1) 首先根据个人知识空间、个人显性信息、个人隐性信息构建个人偏好模型；

2) 基于个人偏好模型，在教材资源、图书资源、试题资源和视音资源的支持下，进行基于偏好、知识点和主题的资源聚合；

3) 然后进行基于偏好的、面向不同用户和不同主题的个性化出版；

4) 接着偏好统计分析；

5) 最后优化个人偏好信息和偏好模型，从而实现后续的个性化出版优化。

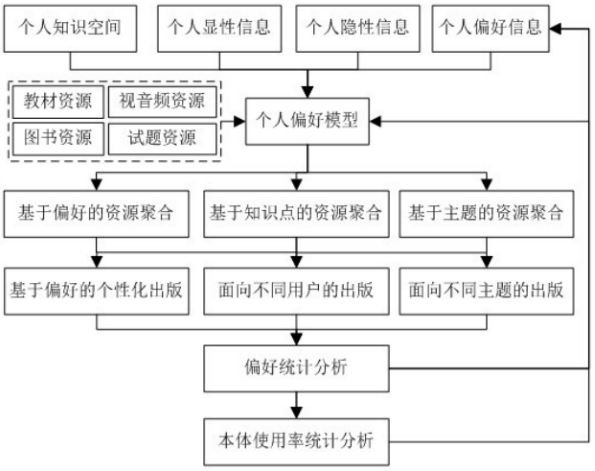

图 2 个人偏好模型知识来源示意图

从图 2 可以看出，主要分为 5 个部分，分别为：个人信息空间，个人知识空间，个人显性信息，个人隐性信息，个人偏好信息。各部分的具体功能为：

1)个人知识空间：个人知识空间记录学生在当前阶段已经掌握的知识或者技能。

2)个人显性信息：显性信息是指用户注册系统时所填写的信息，例如姓名、年级、联系方式等。

3)个人隐性信息：隐性信息是指用户在学习过程中对某个知识点学习频率、学习时间等网络日志的记录以及用户在学习过程中对不同资源类型（如视频、文档等）使用度等。

4)个人偏好信息：通过对用户的显性信息以及隐性信息进行挖掘分析客户的偏好信息。

将图中的个人知识空间、显性信息、隐性信息构成个人的偏好信息，利用偏好信息构建个人偏好模型。个人偏好模型构建的流程如图 3 所示。

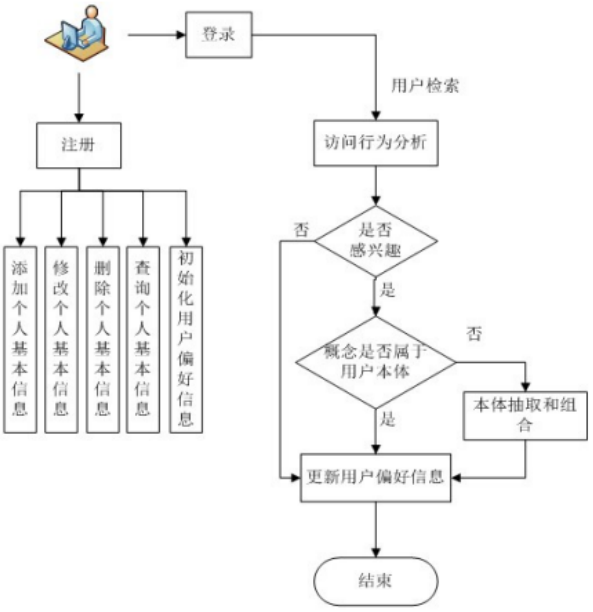

图 3 个人偏好模型构建流程

个人偏好模型构建的流程包括：

1)用户需要创建个人账户，注册个人基本信息，并可以随时修改自己的信息资料，生成个人显性信息，并初始化成用户偏好信息；

2)然后通过网络日志记录用户的学习行为，从而挖掘出用户的隐性信息，用户显性信息与隐性信息构成偏好信息；

3)最后通过用户显性信息与隐性信息的不断变化来更新用户偏好信息，进而形成用户的偏好模型。

4.2 基于个人偏好的资源聚合

在构建了用户偏好模型后，就可以对基于偏好的资源进行聚合，用户信息、偏好、偏号模型、资源库的资源信息的整体聚合过程如图 4 所示。

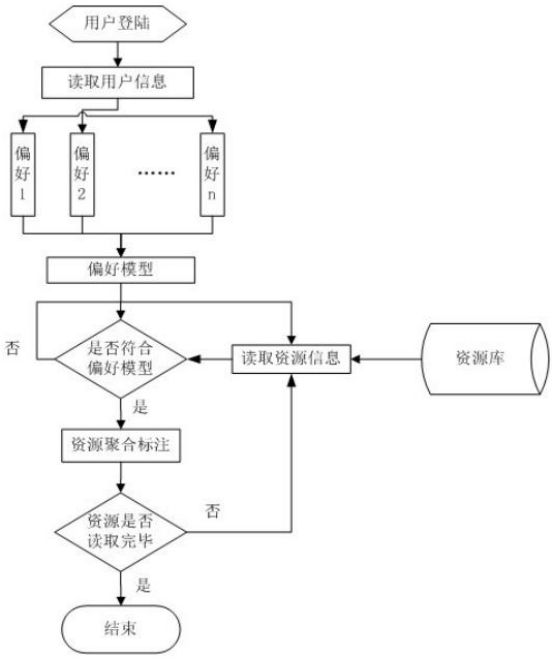

图 4 基于偏好的资源聚合过程示意图

基于偏好的资源整合具体流程为：

1)用户登录后，系统会自动读取用户的偏好模型；

2)根据用户的偏好信息读取已经标注的试题、教材、图书和视音频等资源信息；

3)若资源符合偏好模型则进行资源聚合，如果不是则继续读取资源；

4)当所有的资源都已经读取完毕则结束，此时基于偏好的资源聚合过程完成。

4.3 基于用户偏好的个性化资源推送服务

根据不同用户的偏好模型以推荐相应的资源，以使得用户能够及时有效的获得自己感兴趣的资源。例如用户 A 偏好于视频类资源，用户 B 喜欢文档类的资源，系统就分别推荐相应所偏好的资源，而不是由系统统一的推送同一类资源。

5 结论

本文首先介绍了国内外领域本体知识库的相关理论与方法。接着，具体阐述了构建用户模型和构建领域本体的“兴趣树”的构建方法的相关理论。然后基于已有的 WordNet 语义相似度计算方法，提出了一种基于混合式 WordNeT 的语义相似度计算方法。最后，本文对所构建的用户模型和语义相似度计算方法，用于用户兴趣模型的建立和个性化资源聚合服务，并结合具体实例论证所构建的用户模型和语义计算度的可行性。

参 考 文 献

[1]裴培,丁雪晶.基于本体的语义相似度计算综述[J].合肥学院学报(综合版),2020,37(05):68-74.

[2] 唐晓波,魏巍.基于本体的推荐系统研究综述[J].图书馆学研究,2016(18):7-12.

[3] 李晓光,王大玲,于戈.基于统计语言模型的信息检索 [J].计算机科学,2005,32(08):124－127.

[4]许云,樊孝忠,张锋.基于知网的语义相关度计算.北京理工大学学报,2005,25(05):411－414.

[5]周朴雄,宫楚凡.基于用户动态兴趣标签的推荐模型研究[J].新世纪图书馆,2021(09):65-69.

[6] 荣河江,王亚东.基于基因本体的相似度计算方法[J].智能计算机与应用,2019,9(01):108-113+118.

[7] Gabrilovich E,Markovitch S.Wikipedia-based Semantic Inter-pretation for Natural Language Processing[J]. Journal of Artificial Intelligence Research,2014,34(4):443-498.