(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110347401 B
(45)授权公告日 2021.03.16

(21)申请号 201910527502.2

(22)申请日 2019.06.18

(65)同一申请的已公布的文献号
申请公布号 CN 110347401 A

(43)申请公布日 2019.10.18

(73)专利权人 西安交通大学
地址 710049 陕西省西安市咸宁西路28号

(72)发明人 曲桦 赵季红 边江 张艳鹏 李佳琪 李明霞

(74)专利代理机构 西安通大专利代理有限责任公司 61200
代理人 安彦彦

(51)Int.Cl.
G06F 8/41(2018.01)
G06F 40/30(2020.01)

(56)对比文件
CN 102129479 A,2011.07.20
CN 105404619 A,2016.03.16
CN 106611039 A,2017.05.03
CN 108470181 A,2018.08.31
廉晨思.基于综合本体相似度计算的WEB 服务发现算法.《计算机应用与软件》.2011,第87-89页,转117页.
刘志忠 等.基于语义的服务发现技术研究综述.《计算机工程与科学》.2007,第12-15页,转52页.

审查员 余晓

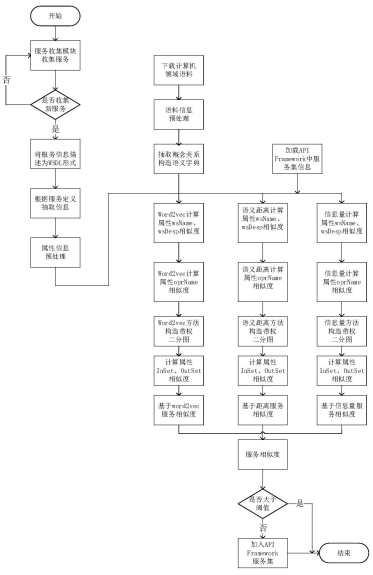
权利要求书3页 说明书10页 附图2页

(54)发明名称

一种基于语义相似度的API Framework服务发现方法

(57)摘要

一种基于语义相似度的API Framework服务发现方法,基于API Framework中服务收集模块不断收集服务信息,将服务信息描述为WSDL标准化文档形式;抽取对应的信息内容构造服务属性信息;将同义词集进行组织,得到两个概念或者两个句子直接的语义相似度;针对属性相似度,分别求解服务之间的相似度,最后获得服务相似度,并判别该服务是否属于新的服务类别,从而判断是否将其加入到API Framework服务集中。本发明通过引入计算机领域的语义词典,通过多种语义相似度计算方法的集成,可以有效的辅助服务信息的发现,增强系统的服务范围。



CN 110347401 B

1.一种基于语义相似度API Framework服务发现方法,其特征在于,包括以下步骤:

1)基于API Framework中服务收集模块不断收集服务信息,如果没有获得服务信息则继续服务收集工作,否则将服务信息描述为WSDL标准化文档形式;同时抽取对应的信息内容构造服务属性信息;

2)将语义词典内所有的术语和概念都以同义词集合的形式表示,将同义词集进行组织,得到知识网络;

3)对步骤2)的知识网络,使用word2vec模型训练分好词的语料,得到训练好的词向量文件,基于训练好的词向量文件和步骤1)得到的服务属性信息,得到两个概念或者两个句子直接的语义相似度;

4)针对属性相似度,分别求解服务之间的相似度,最后使用线性加权的方式获得服务相似度;

5)针对步骤4)得到的服务相似度,判别该服务是否属于新的服务类别,从而做出是否将其加入到API Framework服务集中的决策;

其中,步骤2)中,语义词典通过以下过程构造:选择计算机领域本体的概念及概念关系来源,获取概念及其属性间的关系,生成本体概念层次,并将其映射到OWL语言,采用相关性分析方法对构建的本体网络结构信息进行分析,通过挖掘上下位结构发现不同类目间的关联关系以及进行本体映射研究,并且建立概念的层级,挖掘本体的语义信息,发现本体中的隐含知识,从而构造语义词典;

步骤3)中,两个概念或者两个句子直接的语义相似度$sim_s(t_1,t_2)$采用以下公式计算得到;

$$sim_s(t_1,t_2)=\frac{S_1 \bullet S_2}{\|S_1\| \bullet \|S_2\|} \qquad (1)$$

其中$S_1 \bullet S_2$表示两个句子的向量点乘;$||S_i||$表示句子S_i向量的长度;

步骤5)的具体过程如下:

针对步骤4)得到的服务相似度,根据相似度的判定阈值,如果当前服务相似度大于判定阈值,认为该服务属于系统已有服务,不予加入API Framework服务集中;若当前服务相似度小于判定阈值,则判定其为新的服务类型,将其加入到API Framework服务集中,增加服务的覆盖范围;

步骤4)中,属性相似度通过以下过程得到:针对服务属性信息中的概念,计算概念之间的平均语义相似度;根据概念之间的平均语义相似度得到属性wsName与wsDesp的相似度;属性OprSet中的oprName属性使用字符串匹配算法计算相似度,相等为1,不相等为0;根据两个概念或者两个句子直接的语义相似度、基于距离的语义相似度以及基于信息量的语义相似度,构造带权二分图模型,在该带权二分图模型上计算输入、输出之间的相似度,进而得到属性OprSet中InSet与OutSet属性相似度;

基于距离的语义相似度Sim通过以下过程得到:

$$\mathrm{Sim}=\frac{2*depth(msc(c_1,c_2))}{len(c_1,c_2)+2*depth(msc(c_1,c_2))} \qquad (2)$$

其中,$depth(c_i)$表示概念c_i在语义词典中is_a关系树中的深度,$len(c_1,c_2)$指的是在语

义词典中两个概念(c_1,c_2)最短的路径长度,msc(c_1,c_2)表示概念c_1和概念c_2处于语义词典中is_a关系树中最深层的公共父节点;

在该带权二分图模型上计算输入、输出之间的相似度的具体过程如下:

1)功能性信息语义相似度

语义服务的功能是通过输入和输出属性体现的,接口相似度定义为:

$Sim(ws_i,ws_j)=\alpha \cdot Sim_{In}(ws_i,ws_j)+\beta \cdot Sim_{Out}(ws_i,ws_j)$

其中,$Sim_{In}(ws_i,ws_j)$和$Sim_{Out}(ws_i,ws_j)$分别是服务ws_i和ws_j的输入相似度和输出相似度,α和β是调整因子,且α+β=1;

2)输入相似度

服务的输入是由若干个本体概念组成的,将每个本体概念刻画成一个输入参数,那么就通过一组参数来表示服务的输入;计算2个服务之间的输入相似度是对这2组参数进行匹配;将2组输入参数建模成一个二分图G=($Input_i$,$Input_j$,E),其中$Input_i$和$Input_j$是服务ws_i和ws_j的输入本体概念集合;边集E的构造规则如下:对于$\forall I_i \in Input_i, \forall I_j \in Input_j$,若$Sim_{Concept}(I_i,I_j)>0$,则在二分图G中$I_i$和$I_j$对应的2个节点之间连一条边<$I_i$,$I_j$>,并给该边一个权值$W_{<I_i,I_j>}=Sim_{Concept}(I_i,I_j)$;通过二分图建模之后,参数匹配问题转化为在二分图G上求解集合$Input_i$和$Input_j$的一个最优匹配M,要求最优匹配M的权和最大;根据最优匹配M计算出2个服务间的输入相似度如公式(7)所示;

$$Sim_{In}(ws_i,ws_j)=\frac{\sum_{e\in E_{MIn}} W_e}{|E_{MIn}|} \quad (7)$$

其中E_{MIn}是输入最优匹配MIn的边集;e是MIn中的某一边;W_e是该边的权重;

3)输出相似度

输出相似度计算如公式(8)所示;

$$Sim_{Out}(ws_i,ws_j)=\frac{\sum_{e\in E_{MOut}} W_e}{|E_{MOut}|} \quad (8)$$

其中,E_{MOut}是输出最优匹配MOut的边集,e是MOut中的某一边;W_e是该边的权重。

2.根据权利要求1所述的一种基于语义相似度API Framework服务发现方法,其特征在于:基于信息量的语义相似度通过以下过程得到:

首先计算属于概念节点c中所有单词在语料库中出现的次数freq(c):

$freq(c)=\sum_{n\in words(c)} count(n)$ (3)

其中words(c)表示概念节点c中所包含的所有单词的集合;

概念节点c在语义词典中出现的概率P(c):

$$P(c)=\frac{freq(c)}{Node_{max}} \quad (4)$$

其中$Node_{max}$表示在语义词典中概念节点的总数;

信息量IC(c):

$IC(c)=-\log(P(c))$ (5)

基于信息量计算语义相似度Sim(c_1,c_2)：

$$Sim(c_1,c_2)=\frac{I(common(c_1,c_2))}{I(description(c_1,c_2))}=\frac{2*IC(msc(c_1,c_2))}{IC(c_1)+IC(c_2)} \quad (6)$$

其中，IC(c_i)表示概念节点c_i的信息量，msc(c_1,c_2)表示概念c_1和概念c_2处于语义词典is_a树中最深层的公共父节点。

3.根据权利要求1所述的一种基于语义相似度API Framework服务发现方法，其特征在于：判定阈值为0.8。

一种基于语义相似度的API Framework服务发现方法

技术领域

[0001] 本发明涉及5G网络能力开放应用程序框架(API Framework)的服务发现问题,特别涉及一种基于语义相似度的API Framework服务发现方法。

背景技术

[0002] 工业互联网已成为产业升级发展的必然趋势,5G网络依靠多接入、广覆盖、高性能以及丰富的网络控制和组网方案等优势功能,正在逐渐成为工业企业优先考虑的部署选择。与此同时,5G移动通信系统标准化成为一个亟待解决的问题。在标准化组织第三代合作伙伴计划 (3GPP)中,存在多个北向应用程序接口(API)相关规范(例如,用于3GPP技术规范(TS) 23.682中定义的服务能力暴露功能(SCEF)功能的API,用于MBTS服务提供商和3GPP TR 26.981中定义的BM-SC之间的接口的API)。为避免不同API规范之间的方法重复和不一致,3GPP考虑开发通用API框架(CAPIF),其中包括适用于任何北向服务API的常见方面。

[0003] 参照R15能力开放标准规范,研发出API网关(GW)系统,提供基于Restful的网络能力开放API,为第三方提供调用获得相关网络能力(如用户位置、基于业务要求的数据传输QoS保障等)的服务;设计能力开放API Framework,实现API的注册,发现及授权机制,进行标准化操作从而弥补现有技术中不能为API Framework的管理提供一套完整解决方案的问题。

[0004] 基于5G通信技术和网络的不断发展和演进,在API Framework中构造Restful形式的服务集,为用户提供更加高效、便捷的服务。然而在服务大规模增长的场景下,API Framework 如何在保证服务集完备性和健壮性的前提下,通过服务之间相似度的衡量,从而发现新的服务类别来扩展该API Framework的服务范围。其中,服务发现依赖于服务之间的相似度,而服务之间的相似度最终转化为对于API中的概念语义相似度的计算,传统的语义相似度计算方法分为四类,分别为基于距离的方法、基于信息量的方法、基于属性的方法以及混合式方法。这四种语义相似度计算方法特点如下:

[0005] (1)基于距离的方法

[0006] 基于距离计算语义相似度的思想:在本体概念结构树中通过向量化概念词来计算两个概念之间的路径长度,通过两个概念在路径维度上的结果得出相似度关系,规定两个概念词在本体层次树中的路径长度越大,相似度越小。传统基于距离的计算方法主要利用了语义字典 (WordNet)中的上下位结构信息来计算相似度,方法简单易实施,但是由于只利用距离、深度、宽度等语义信息进行计算,在计算的准确性方面表现较差,从而影响了服务发现的效率。

[0007] (2)基于信息量的方法

[0008] 基于信息量计算语义相似度的方法是将概念的信息量与本体知识相结合,即认为概念对之间共享信息量越高,其概念的差异信息量越少,相似度则越高。其中,共享信息量根据共享的父节点信息量计算,而差异信息量根据各个概念与共享父节点的差量来计算。在本体的概念结构树中,每个概念子节点可以认为是根节点的实例化以及概念的扩展,因

此根据其父节点之间的信息量的关系可以计算其概念之间的相似度。基于信息量的语义计算方法能够客观的反应概念节点在语义、语法等方面的相似度和差异性，但其最大的问题是，信息量的计算依赖于语料库，不同的语料库存在较大的差异，使用不同的语料库计算也会产生很大的差异，从而导致其语义相似度的计算很难形成统一的结果，从而影响其服务发现的可信度。

[0009] (3)基于属性的方法

[0010] 基于属性的语义相似度计算方法主是通过两个概念之间属性集的相似程度来衡量语义的相似度关系。基于属性的语义相似度基于本体属性的重叠程度来计算语义相似度，从而更好地解决跨本体的语义相似度问题，从而很好的弥补了基于距离计算语义相似度时无法跨本体的问题。基于属性的方法依赖概念节点具有完备的属性集，对于WordNet等大型本体字典才会拥有丰富的语义知识，而其他的特定领域词典不会含有足够的语义内容，从而造成属性的相似度无法有效的计算，影响最终语义相似度的准确率，导致其服务发现的效率比较低。

[0011] (4)混合式方法

[0012] 不同于基于距离、基于信息量以及基于属性的计算方法，混合式的计算方法充分利用多种语义信息，对于各种计算因素加以不同的权重，得到最终的相似度计算结构。由于使用语义信息更加充分，该方法能够极大的挖掘语义信息来提高准确度。然而，由于其需要根据领域本体设置权值，从而权重设置的不确定性影响了这种方法的普适性，通用的语义词典很难满足特定的业务需求，导致在设定的应用场景下服务发现效率比较低。

[0013] 考虑到传统方法严重影响服务发现效率的情况，因而，计算方法并不能直接应用于API Framework服务发现中，必须根据其所属领域对其相似度计算方法进行改进。

发明内容

[0014] 本发明的目的在于解决API Framework系统中服务发现问题，提供一种基于语义相似度的API Framework服务发现方法，该方法在构建基于计算机领域知识网络的基础上，通过计算服务之间各个属性的相似度，从而对服务之间的总体相似度进行准确的分析，根据分析的结果做出决策，可以有效的解决由于语义相似度计算准确性不够而引起无法准确发现服务的情况。

[0015] 为了达到上述目的，本发明采用了以下技术方案：

[0016] 一种基于语义相似度API Framework服务发现方法，包括以下步骤：

[0017] 1)基于API Framework中服务收集模块不断收集服务信息，如果没有获得服务信息则继续服务收集工作，否则将服务信息描述为WSDL标准化文档形式；同时抽取对应的信息内容构造服务属性信息；

[0018] 2)将语义词典内所有的术语和概念都以同义词集合的形式表示，将同义词集进行组织，得到知识网络；

[0019] 3)对步骤2)的知识网络，使用word2vec模型训练分好词的语料，得到训练好的词向量文件，基于训练好的的词向量文件和步骤1)得到的服务属性信息，得到两个概念或者两个句子直接的语义相似度；

[0020] 4)针对属性相似度，分别求解服务之间的相似度，最后使用线性加权的方式获得

服务相似度；

[0021] 5)针对步骤4)得到的服务相似度，判别该服务是否属于新的服务类别，从而做出是否将其加入到API Framework服务集中的决策。

[0022] 本发明进一步的改进在于：步骤2)中，语义词典通过以下过程构造：选择计算机领域本体的概念及概念关系来源，获取概念及其属性间的关系，生成本体概念层次，并将其映射到OWL语言，采用相关性分析方法对构建的本体网络结构信息进行分析，通过挖掘上下位结构发现不同类目间的关联关系以及进行本体映射研究，并且建立概念的层级，挖掘本体的语义信息，发现本体中的隐含知识，从而构造语义词典。

[0023] 本发明进一步的改进在于：步骤3)中，两个概念或者两个句子直接的语义相似度$sim_s(t_1,t_2)$采用以下公式计算得到；

[0024] $$sim_s(t_1,t_2)=\frac{S_1\bullet S_2}{\|S_1\|\bullet\|S_2\|} \quad (1)$$

[0025] 其中$S_1\bullet S_2$表示两个句子的向量点乘；$||S_i||$表示句子S_i向量的长度。

[0026] 本发明进一步的改进在于：步骤4)中，属性相似度通过以下过程得到：针对服务属性信息中的概念，计算概念之间的平均语义相似度；根据概念之间的平均语义相似度得到属性 wsName与wsDesp的相似度；属性OprSet中的oprName属性使用字符串匹配算法计算相似度，相等为1，不相等为0；根据两个概念或者两个句子直接的语义相似度、基于距离的语义相似度以及基于信息量的语义相似度，构造带权二分图模型，在该带权二分图模型上计算输入、输出之间的相似度，进而得到属性OprSet中InSet与OutSet属性相似度。

[0027] 本发明进一步的改进在于：基于距离的语义相似度Sim通过以下过程得到：

[0028] $$\mathrm{Sim}=\frac{2*depth(msc(c_1,c_2))}{len(c_1,c_2)+2*depth(msc(c_1,c_2))} \quad (2)$$

[0029] 其中，$depth(c_i)$表示概念c_i在语义词典中is_a关系树中的深度，$len(c_1,c_2)$指的是在语义词典中两个概念(c_1,c_2)最短的路径长度，$msc(c_1,c_2)$表示概念c_1和概念c_2处于语义词典中is_a关系树中最深层的公共父节点。

[0030] 本发明进一步的改进在于：基于信息量的语义相似度通过以下过程得到：

[0031] 首先计算属于概念节点c中所有单词在语料库中出现的次数freq(c)：

[0032] $\mathrm{freq}(c)=\sum_{n\in words(c)}\mathrm{count}(n)$ (3)

[0033] 其中words(c)表示概念节点c中所包含的所有单词的集合；

[0034] 概念节点c在语义词典中出现的概率P(c)：

[0035] $$P(c)=\frac{freq(c)}{Node_{max}} \quad (4)$$

[0036] 其中$Node_{max}$表示在语义词典中概念节点的总数；

[0037] 信息量IC(c)：

[0038] $\mathrm{IC}(c)=-\log(P(c))$ (5)

[0039] 基于信息量计算语义相似度$Sim(c_1,c_2)$：

[0040] $$Sim(c_1,c_2)=\frac{I(common(c_1,c_2))}{I(description(c_1,c_2))}=\frac{2*IC(msc(c_1,c_2))}{IC(c_1)+IC(c_2)} \quad (6)$$

[0041] 其中，IC(c_i)表示概念节点c_i的信息量，msc(c_1,c_2)表示概念c_1和概念c_2处于语义词典is_a树中最深层的公共父节点。

[0042] 本发明进一步的改进在于：步骤5)的具体过程如下：

[0043] 针对步骤4)得到的服务相似度，根据相似度的判定阈值，如果当前服务相似度大于判定阈值，认为该服务属于系统已有服务，不予加入API Framework服务集中；若当前服务相似度小于判定阈值，则判定其为新的服务类型，将其加入到API Framework服务集中，增加服务的覆盖范围。

[0044] 本发明进一步的改进在于：判定阈值为0.8。

[0045] 与现有技术相比，本发明具有以下有益效果：本发明引入基于计算机领域知识网络的语料信息，对所发现的服务与系统中服务集合中的服务进行相似度计算，可以有效的提高服务的发现能力，维护API Framework的服务集完备性。本发明考虑到基于WordNet的词汇语义相似度计算方法中隔离抽象词汇和具象词汇，以及片面依赖上下文关系的不足，提出了基于计算机领域知识网络的概念语义相似度计算方法。同时，考虑到传统中基于WordNet语义词典计算语义相似度存在准确性不够、严重依赖语料库等缺点，提出将基于Word2vec的语义计算方法、基于距离的语义计算方法以及基于信息量的计算方法进行集成的方法，从而保证其在准确性方面的优势。通过改进语料词典和相似度的计算方法，可以很好的弥补传统方法中存在的不足(如服务发现效率比较低、不能适应多种应用场景等)，同时针对其特定的应用场景，保证服务发现的高效性。

附图说明

[0046] 图1为基于能力开放API Framework中服务发现模块图。

[0047] 图2为基于服务信息的数据预处理流程图。

[0048] 图3为基于带权二分图模型计算接口相似度流程图。

[0049] 图4为基于语义相似度在服务发现中决策流程图。

具体实施方式

[0050] 为了使本发明的内容、效果以及优点更加清楚明白，下面结合附图和实施例对本发明进行详细描述。

[0051] 本发明是应用构造基于计算机领域的语料知识，基于集成的语义相似度策略计算服务之间的相似度，通过对API Framework中服务收集模块收集的服务信息进行服务判定，在保证语义相似度的计算准确性的前提下，可以有效的避免传统语义计算方法中准确率不高、严重依赖领域词典的缺点，可以有效的提高系统中服务发现的能力，从而提高系统服务集合的服务覆盖能力。参见图1，图1中展示了基于能力开放API Framework中服务发现模块，本发明中服务来源于API Framework中服务发现模块，通过对于发现的服务与系统服务集合中服务之间的相似度进行计算，判别该服务是否作为新的服务类型加入到服务集合，在避免大量人工参与的情况下，保证提高服务的支持能力，致力于为用户提供高效的服务体验。

[0052] 本发明的具体过程如下：

[0053] (一)服务相关定义

[0054] 定义1:网络能力开放服务是一种可以通过网络通信的应用程序,通过标准的网络协议提供服务,目的保证不同平台的应用服务可以互操作,通常表现为一个向外界暴露出能够通过Internet进行调用的API。

[0055] 定义2:WSDL是能力开放服务的描述语言,以一种基于XML语言描述服务的文件形式,描述了调用服务所需要的详细信息,它描述说明三个基本属性:

[0056] 服务做些什么:服务所提供的操作方法

[0057] 如何访问服务:数据格式详情以及访问服务操作的必要协议

[0058] 服务位于何处:有特定的协议决定的网络地址

[0059] 能力开放服务的描述信息是服务相似度计算的基础,服务的语义描述可以抽象定义为如下:

[0060] 定义3:能力开放服务:能力开放服务的描述表示为一个五元组ws＝{wsId, wsName, wsDesp,OprSet,wsAddr},其中:

[0061] (1)wsId是服务的编号,每个服务在服务集合中的唯一标识符;

[0062] (2)wsName是服务的名称信息;

[0063] (3)wsDesp是服务功能的详细文本描述;

[0064] (4)OprSet是服务操作集合,OprSet＝{opr_1,opr_2,..opr_n},其中opr_i表示一个服务操作;

[0065] (5)wsAddr是服务请求访问的地址;

[0066] 每个服务操作对应一组服务操作方法、使用该方法的服务的输入接口信息以及调用该方法获得的输出接口信息,结合参数之间的关联关系,服务操作定义如下:

[0067] 定义4:服务操作:服务操作表示为一个三元组opr＝{oprName,InSet,OutSet},其中:

[0068] (1)oprName是服务操作的名称;

[0069] (2)InSet＝{inP_1,inP_2,...inP_n}是输入接口信息集合,其中inP_i为第i个输入接口信息,i取值为1～n;

[0070] (3)OutSet＝{$outP_1$,$outP_2$,...$outP_n$}是输出接口信息集合,其中$outP_i$为第i个输出接口信息, i取值为1～n;

[0071] (二)基于能力开放API framework服务发现

[0072] 参照R15能力开放标准规范,设计基于Restful的网络能力开放API Framework,其提供了各种模块来应用API,其中包括API发现、API注册、API授权、API安全等模块,本发明基于语义相似度来完成服务的发现功能,从网络中获得5G相关服务的过程主要包括以下步骤:

[0073] 1)服务提供商开发出新的服务,将服务注册请求提交到UDDI(Universal Description, Discovery and Integration,即通用描述、发现与集成服务),UDDI对Web Service(服务注册请求)进行审核,当审核通过时,同意将新服务注册到UDDI服务目录中;

[0074] 2)API Framework中提供的服务收集模块定期从UDDI服务目录中搜索服务,服务收集模块维护一个服务标记文件,标记每一个服务是否被访问过,当服务收集模块从UDDI中收集服务且该服务没有被访问过时,获取关于该服务的所有描述信息;

[0075] 3)服务收集模块将获取的描述信息转化为标准的WSDL服务描述文件,应用本发明

所设计的API Framework服务发现方法,进而维护服务之间的关系。

[0076] (三)数据预处理与语义词典构造

[0077] 1)数据预处理

[0078] WSDL是一个描述服务信息的XML文档,包含7个重要元素,分别types,import,message, portType,operation,binding,service。根据WSDL文档特点,对于步骤(二)中服务发现模块发现的服务信息,构造其对应的信息内容,实现服务信息的规范化处理,同时为了完成服务的相似度计算,需要将每一个服务的详细描述文本信息转化为上文中服务定义的格式,从而支撑下文中的相似度计算。

[0079] 参见图2,根据自然语言处理的主要流程,对WSDL格式服务文件进行信息提取,得到 Web信息,然后进行分词,加载停用词表,去停用词、去标点后大写还原,词干化,在进行词性标注,最后输出标准化服务信息。服务信息的数据预处理主要分为以下步骤:

[0080] 1.1)分词:基于宾夕法尼亚大学计算机和信息科学使用python语言实现的自然语言工具包NLTK,对于一个服务中各个属性的描述信息,调用其分词的方法实现分词的效果;

[0081] 1.2)去停用词:加载维基百科提供的停用词表,实现对于停用词的过滤作用。

[0082] 1.3)词干化:对于词干化处理,同样适用NLTK工具包中提供的词干化功能实现。

[0083] 1.4)词性标注:使用NLTK中的词性标注模块进行处理,给出每个词不同的词性。

[0084] 2)语义词典构造

[0085] 针对传统基于WordNet的词语语义相似度计算方法中隔离抽象词汇和具象词汇以及片面依赖上下文关系的缺点,针对本发明研究的API Framework中服务发现的背景,提出基于计算机领域知识网络的词汇语义相似度计算方法。同时基于上下文、工具-工具对象、部件-整体等概念关系准则构建计算机词汇的知识网络,提出集成多种语义相似度的计算方法来弥补单个方法中存在的不足,得到更高的语义一致性结果,提高服务发现的能力。

[0086] 本发明选择计算机领域本体的概念及概念关系来源,获取概念及其属性间的关系,生成本体概念层次,并将其映射到OWL语言,采用相关性分析方法对构建的本体网络结构信息进行分析,可以通过挖掘上下位结构发现不同类目间的关联关系以及进行本体映射研究,并且可以利用这种形式框架辅助建立概念的层级,以期充分挖掘本体的语义信息,发现本体中的隐含知识,从而构造语义词典。

[0087] (四)基于Word2vec计算语义相似度

[0088] Word2vec模型是一款将词汇向量化的高效工具,其思想是经过训练将每个词汇映射成K 维实数向量(K为模型中的超参数),通过计算词汇之间的距离(如欧式距离等)来判断它们的相似程度。Word2vec中包含两个模型分别为CBOW模型和Skip-gram模型,其中CBOW为连续词袋模型,即利用词汇的上下文来预测该词汇,而Skip-gram将当前中心词作为输入,预测上下文信息。

[0089] 基于下载的计算机领域的语料,对语料进行基本的预处理操作,word2vec模型在给定的语料库上训练CBOW和Skip-Gram两种模型,应用服务定义得到服务的概念描述信息,在已经训练好的Word2vec模型上计算得到所有概念在语料库上的词向量表示。

[0090] 针对所有概念在语料库上的词向量表示,使用余弦距离计算两个概念或者两个句子直接的语义相似度,依照公式(1)计算语义相似度。

[0091] $$sim_s(t_1,t_2)=\frac{S_1\bullet S_2}{\|S_1\|\bullet\|S_2\|} \quad (1)$$

[0092] 其中$S_1 \bullet S_2$表示两个句子的向量点乘；||S_i||表示句子S_i向量的长度；

[0093] (五)基于距离计算语义相似度

[0094] 对于定义3中的能力开放服务属性描述信息，采用步骤(三)中的数据预处理方法进行处理，获得需要判断服务的属性概念信息。通过路径距离的语义相似度算法以语义词典中的“is_a”关系分类树为基础，通过两个概念在关系树中的最短路径来表示它们之间的语义相似度，该类方法认为距离越近的概念间语义相似度越高。

[0095] 为了充分利用概念的语义中的各种信息，在考虑两个概念在关系树中的路径长度时，也考虑最小父节点在关系树中的深度。当两个节点长度相同时，它们的父节点越深，则相似度越大，而父节点相同的节点，长度越大，相似度越低。依据公式(2)计算基于距离的语义相似度Sim。

[0096] $$\mathrm{Sim}=\frac{2*depth(msc(c_1,c_2))}{len(c_1,c_2)+2*depth(msc(c_1,c_2))} \quad (2)$$

[0097] 其中，depth(c_i)表示概念c_i在语义词典中“is_a”关系树中的深度。len(c_1,c_2)指的是在语义词典中两个概念(c_1,c_2)最短的路径长度，msc(c_1,c_2)表示概念c_1和概念c_2处于语义词典中“is_a”关系树中最深层的公共父节点。

[0098] (六)基于信息量计算语义相似度

[0099] 对于定义3中的能力开放服务属性描述信息，采用步骤(三)中的数据预处理方法进行处理。在语义词典的树形结构中，每个概念节点的子节点都是对其祖先节点所表达概念的一次细分和具体化，因此，可以通过被比较概念节点对其祖先节点所包含的信息量来衡量它们直接的相似度，所以在比较概念节点的相似度之前，需要计算语义词典中各个节点的信息量。

[0100] 首先计算属于概念节点c中所有单词在语料库中出现的次数，次数freq(c)统计公式如式 (3)所示。

[0101] $$\mathrm{freq}(c)=\sum_{n\in words(c)}\mathrm{count}(n) \quad (3)$$

[0102] 其中words(c)表示概念节点c中所包含的所有单词的集合；

[0103] 概念节点c在语义词典中出现的概率，概率P(c)计算如式(4)所示。

[0104] $$P(c)=\frac{freq(c)}{Node_{max}} \quad (4)$$

[0105] 其中$Node_{max}$表示在语义词典中概念节点的总数；

[0106] 信息量(IC)的计算如公式(5)所示。

[0107] $$\mathrm{IC}(c)=-\log(P(c)) \quad (5)$$

[0108] 任意一对概念的相似度和它们的共性相关，共性越大，相似度越高。基于信息理论的定义和信息量，给出基于信息量计算语义相似度的计算如公式(6)所示。

[0109] $$Sim(c_1,c_2)=\frac{I(common(c_1,c_2))}{I(description(c_1,c_2))}=\frac{2*IC(msc(c_1,c_2))}{IC(c_1)+IC(c_2)} \quad (6)$$

[0110] 其中，IC(c_i)表示概念节点c_i的信息量，msc(c_1,c_2)表示概念c_1和概念c_2处于语义

词典“is_a”树中最深层的公共父节点。

[0111] (七)基于带权二分图模型计算接口相似度

[0112] 计算出概念之间的语义相似度后,就可以构建不同服务输入接口间带权二分图和输出接口间带权二分图,概念节点间的语义相似度作为权重。这样,服务间的功能性语义信息的匹配,即服务的输入和输出接口的匹配转化为计算带权二分图中的相似程度(图3给出基于带权二分图模型计算接口相似度流程图)。参见图3,基于带权二分图模型计算接口相似度的具体过程如下:

[0113] 1)功能性信息语义相似度

[0114] 语义服务的功能是通过输入和输出属性体现的,接口相似度定义为:

[0115] $Sim(ws_i,ws_j)=\alpha\cdot Sim_{In}(ws_i,ws_j)+\beta\cdot Sim_{Out}(ws_i,ws_j)$

[0116] 其中,$Sim_{In}(ws_i,ws_j)$和$Sim_{Out}(ws_i,ws_j)$分别是服务ws_i和ws_j的输入相似度和输出相似度,α和β是调整因子,且α+β=1;

[0117] 2)输入相似度

[0118] 根据定义4可知,服务的输入是由若干个本体概念组成的,可将每个本体概念刻画成一个输入参数,那么就可以通过一组参数来表示服务的输入。计算2个服务之间的输入相似度就是对这2组参数进行匹配。为了解决这类参数的匹配问题,可以把2组输入参数建模成一个二分图$G=(Input_i,Input_j,E)$,其中$Input_i$和$Input_j$是服务ws_i和ws_j的输入本体概念集合;边集E的构造规则如下:对于$\forall I_i\in Input_i,\forall I_j\in Input_j$,若$Sim_{Concept}(I_i,I_j)>0$,则在二分图G 中$I_i$和$I_j$对应的2个节点之间连一条边$\langle I_i,I_j\rangle$,并给该边一个权值$W_{\langle Ii,Ij\rangle}=Sim_{COncept}(I_i,I_j)$。通过二分图建模之后,参数匹配问题就可以转化为在二分图G上求解集合$Input_i$和$Input_j$的一个最优匹配M,要求最优匹配M的权和最大。根据最优匹配M可以计算出2个服务间的输入相似度如公式(7)所示。

[0119] $$Sim_{In}(ws_i,ws_j)=\frac{\sum_{e\in E_{MIn}}W_e}{|E_{MIn}|} \quad (7)$$

[0120] 其中E_{MIn}是输入最优匹配MIn的边集;e是MIn中的某一边;W_e是该边的权重;

[0121] 3)输出相似度

[0122] 与输入相似度类似,输出相似度计算如公式(8)所示。

[0123] $$Sim_{Out}(ws_i,ws_j)=\frac{\sum_{e\in E_{MOut}}W_e}{|E_{MOut}|} \quad (8)$$

[0124] 其中,E_{MOut}是输出最优匹配MOut的边集,e是MOut中的某一边;W_e是该边的权重。

[0125] (八)计算服务之间的相似度

[0126] 参见图4,服务之间的相似度计算流程(图4给出基于语义相似度在服务发现中决策流程图)如下:

[0127] 1)基于API Framework中服务收集模块不断收集服务信息,如果没有获得服务信息则继续服务收集工作,否则进行服务信息描述相关步骤,即将其服务信息描述为WSDL标准化文档形式;同时根据服务的定义3的能力开放服务和定义4的服务操作,抽取对应的信息内容构造服务属性信息;

[0128] 2)利用步骤(三)构造的语义词典进行知识网络的构建,即将语义词典内所有的术语和概念都以同义词集合的形式表示,针对每一个同义词集都有一个简单的定义描述以及该同义词集存在的语义关系记录;将同义词集按照上下位、整体、部分、同义、反义、因果等关系组织起来,得到知识网络,为每一个术语提供严格的语义层次结构;

[0129] 3)针对服务中描述的属性信息分别进行一系列的预处理操作,包括分词、去停用词、词干化、词性标注等基本数据处理逻辑,从而得到针对每个属性的一组描述信息;即采用步骤 (三)中的数据预处理过程进行处理。

[0130] 4)对步骤2)构建的知识网络,使用word2vec模型训练分好词的语料,得到训练好的词向量文件,基于训练好的的词向量文件和步骤1)得到的服务属性信息,根据步骤(四)中的方法,得到两个概念或者两个句子直接的语义相似度;

[0131] 5)基于步骤2)中构造的语义词典,应用步骤(五)中的方法,得到基于距离的语义相似度;

[0132] 6)基于步骤2)中构造的语义词典,应用步骤(六)中的方法,得到基于信息量的语义相似度;

[0133] 7)根据定义3中的服务属性之间的特征,分别制定不同的策略来计算对应的相似度,其中属性wsName,wsDesp由简单的描述信息构成,针对服务属性中的概念,计算概念之间的平均语义相似度;根据概念之间的平均语义相似度得到属性wsName,wsDesp的相似度;由于Restful风格服务中的方法主要为GET、PUT、POST、DELETE四种方法,所以属性OprSet中的oprName属性使用字符串匹配算法计算相似度,相等为1,不相等为0;根据步骤4),步骤5)以及步骤6)中求解概念相似度的方法,构造带权二分图模型,在该带权二分图模型上计算输入、输出之间的相似度,进而得到属性OprSet中InSet,OutSet属性相似度;针对步骤3)的属性描述信息,应用步骤4),步骤5)以及步骤6)提供的三种不同求解概念相似度的方法,从而得到每个属性的不同相似度;

[0134] 8)针对步骤7)中得到的每个属性的不同相似度,分别求解服务之间的相似度,最后使用线性加权的方式获得服务相似度。

[0135] (九)决策

[0136] 针对步骤(八)得到的服务相似度,判别该服务是否属于新的服务类别,从而做出是否将其加入到API Framework服务集中的决策。主要判别步骤如下:

[0137] 对于计算得到的服务相似度,根据相似度的判定阈值(在本发明设定判定阈值为0.8),如果当前服务相似度大于判定阈值,认为该服务属于系统已有服务,不予加入API Framework 服务集中;若当前服务相似度小于判定阈值,则判定其为新的服务类型,将其加入到API Framework服务集中,增加服务的覆盖范围。

[0138] 本发明设计了一种基于语义相似度的网络服务发现机制,以提升服务发现的有效性和实时性。本发明在API Framework的服务发现问题上,应用自然语言处理的基础手段,对获取的服务详细信息解析以获得服务信息的标准化描述,应用基本的语言处理逻辑,加强语义相似度计算的容错性;同时本发明通过引入基于计算机领域的知识网络,发现传统语义计算方法存在的问题,提出基于Word2vec的语义计算方法、基于距离的语义计算方法以及基于信息量的语义计算方法的集成,使得基于语义计算两个服务之间的相似度拥有很高的准确率,提升了服务发现的效率。

[0139] 本发明通过引入语义相似度计算模型，对API Framework服务信息进行发现，可以准确的判别服务与现有API Framework服务集的关系，从而有效扩展服务集，提高系统的服务能力。通过构造基于计算机领域的语义词典，可以准确的描述概念的语义，避免由于词典语义信息不足而导致概念相似度无法衡量的问题；通过基于word2vec的概念计算方法、基于距离的概念计算方法以及基于信息量的概念计算方法的集成，保证在充分应用语义信息的同时，有效提高服务之间相似度计算的准确性，避免由于单一方法引起的语义信息描述不准确的问题。本发明通过引入计算机领域的语义词典，通过多种语义相似度计算方法的集成，可以有效的辅助服务信息的发现，增强系统的服务范围。

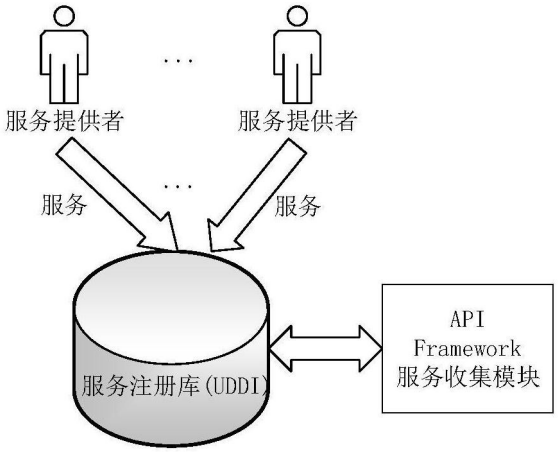

服务提供者
服务提供者
服务
服务
服务注册库(UDDI)
API
Framework
服务收集模块

图1

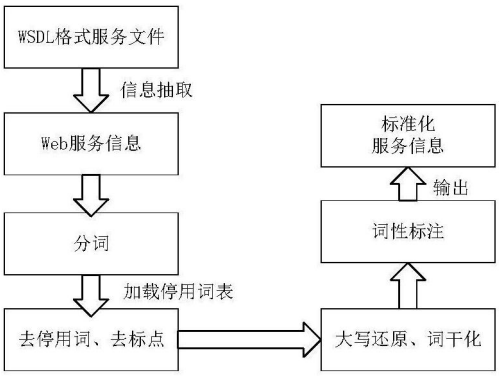

WSDL格式服务文件
信息抽取
Web服务信息
分词
加载停用词表
去停用词、去标点
大写还原、词干化
词性标注
输出
标准化
服务信息

图2

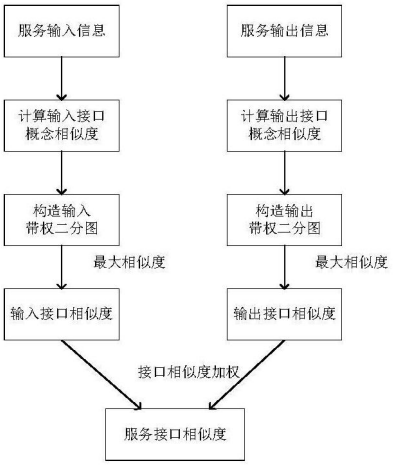

服务输入信息
服务输出信息
计算输入接口
概念相似度
计算输出接口
概念相似度
构造输入
带权二分图
构造输出
带权二分图
最大相似度
最大相似度
输入接口相似度
输出接口相似度
接口相似度加权
服务接口相似度

图3

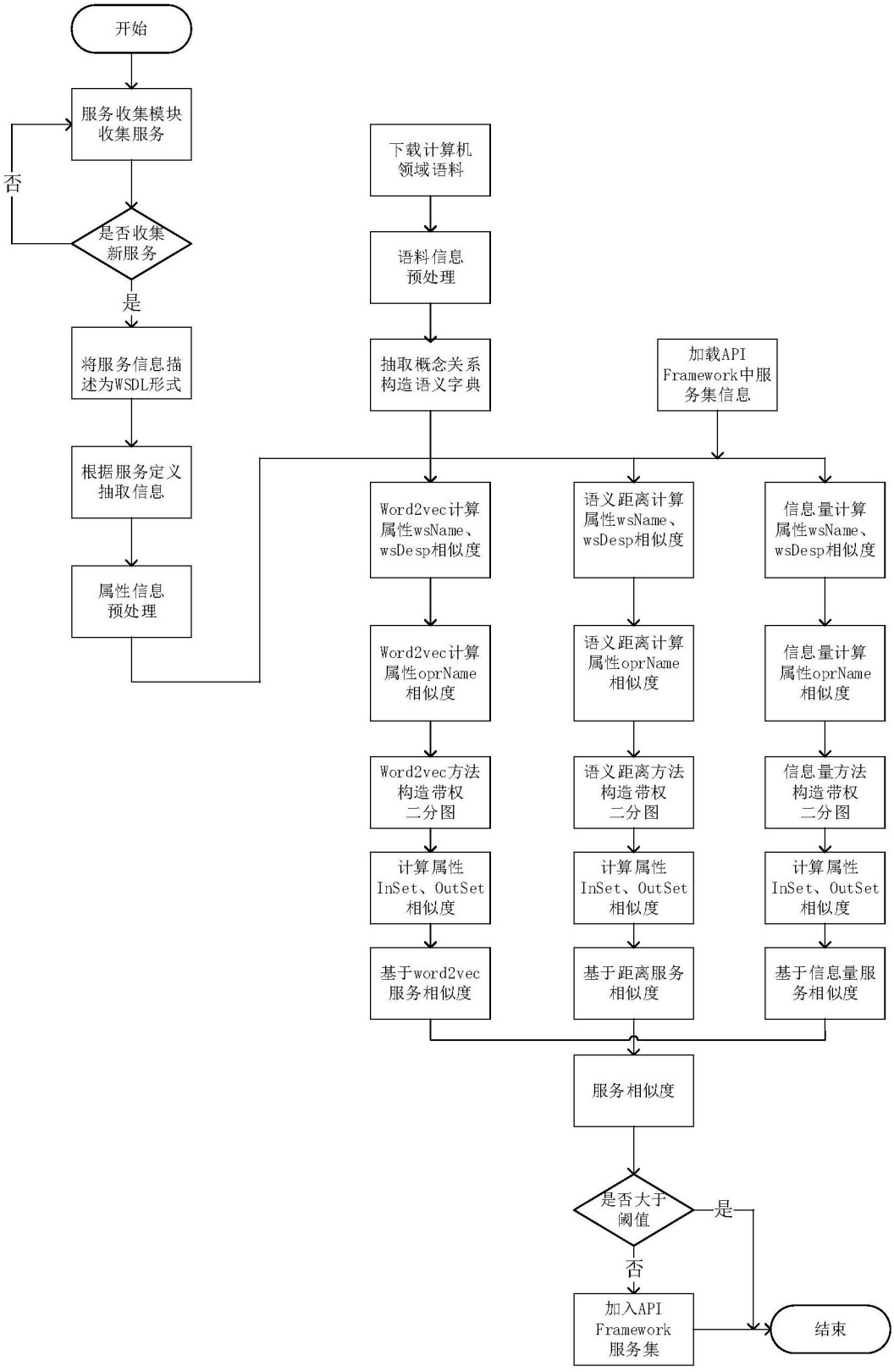

开始
服务收集模块收集服务
否
是否收集新服务
是
将服务信息描述为WSDL形式
根据服务定义抽取信息
属性信息预处理
下载计算机领域语料
语料信息预处理
抽取概念关系构造语义字典
加载API Framework中服务集信息
Word2vec计算属性wsName、wsDesp相似度
语义距离计算属性wsName、wsDesp相似度
信息量计算属性wsName、wsDesp相似度
Word2vec计算属性oprName相似度
语义距离计算属性oprName相似度
信息量计算属性oprName相似度
Word2vec方法构造带权二分图
语义距离方法构造带权二分图
信息量方法构造带权二分图
计算属性InSet、OutSet相似度
计算属性InSet、OutSet相似度
计算属性InSet、OutSet相似度
基于word2vec服务相似度
基于距离服务相似度
基于信息量服务相似度
服务相似度
是否大于阈值
是
否
加入API Framework服务集
结束

图4