

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2019 年 10 月 24 日 (24.10.2019)



(10) 国际公布号
WO 2019/201068 A1

(51) 国际专利分类号:
H04L 29/08 (2006.01)

(21) 国际申请号: PCT/CN2019/079967

(22) 国际申请日: 2019 年 3 月 28 日 (28.03.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810345829.3 2018年4月18日 (18.04.2018) CN

(71) 申请人: 山东科技大学(SHANDONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) [CN/CN]; 中国山东省青岛市黄岛区前湾港路 579 号, Shandong 266590 (CN)。

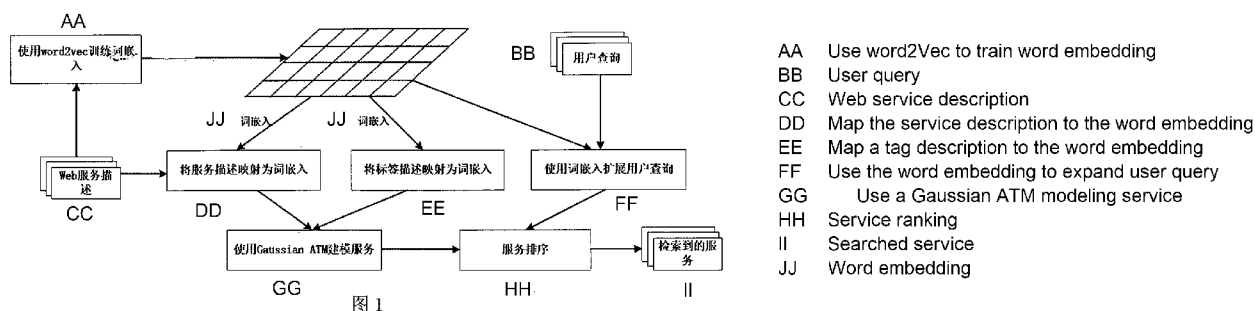
(72) 发明人: 田刚(TIAN, Gang); 中国山东省青岛市黄岛区前湾港路 579 号, Shandong 266590 (CN)。刘鹏飞(LIU, Pengfei); 中国山东省青岛市黄岛区前湾港路 579 号, Shandong 266590 (CN)。孙承爱(SUN, Chengai); 中国山东省青岛市黄岛区前湾港路 579 号, Shandong 266590 (CN)。

(74) 代理人: 青岛智地领创专利代理有限公司(QINGDAO ZHIDILINGCHUANG PATENT AGENCY CO.,LTD); 中国山东省青岛市黄岛区江山路 480 号青岛研创中心 4 号楼 602 室雷钰, Shandong 266555 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) Title: SEMANTIC SPARSE WEB SERVICE DISCOVERY METHOD BASED ON GAUSSIAN ATM AND WORD EMBEDDING

(54) 发明名称: 一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法



(57) Abstract: The present invention provides a semantic sparse web service discovery method based on Gaussian ATM and word embedding. The method is performed in the following order: collecting a Web service description document, parsing the collected Web service description document, obtaining a feature vocabulary in the Web service description document, and preprocessing the vocabulary in the Web service description document to obtain a set of prototype words; using a word embedding training model Word2Vec to train the set of prototype words to obtain a continuous vector representation of each word in the set of prototype words; using the set of prototype words obtained by the Gaussian ATM model to train, and obtaining each Web service hierarchical structure; using the set of continuous vectors obtained by training to enrich user query, and obtaining an extended user query; using the obtained service hierarchical structure and using a probability ranking method to obtain query output corresponding to the extended user query. The method provided in the present invention can implement accurate and efficient Web service discovery.

(57) 摘要: 本发明提供了一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法, 按如下顺序进行: 收集Web服务描述文档, 对收集的Web服务描述文档进行解析, 得到Web服务描述文档中的特征词汇, 并对Web服务描述文档中的词汇进行预处理, 得到一组原型单词的集合; 使用词嵌入训练模型Word2Vec训练原型单词的集合, 得到原型单词的集合中每个单词的连续向量表示; 使用Gaussian ATM模型得到的原型单词的集合进行训练, 获得每个Web服务层次结构; 使用训练得到的连续向量的集合丰富用户查询, 得到扩展以后的用户查询; 利用得到的服务层次结构, 使用概率排序方法, 得到针对扩展之后用户查询所对应的查询输出。本发明提供的方法能实现准确、高效的Web服务发现。

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法

技术领域

[0001] 本发明涉及服务计算技术领域，具体涉及一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法。

背景技术

[0002] 在信息化日益普及的今天，互联网的软件体系结构和其计算环境发生着急剧的变化，促成了基于互联网体系的软件服务开发的发展和普及。在多变的软件服务环境下，由于各种计算资源具有高度分布和自治的根本特征，所以软件系统的开发将会面临新挑战。可以使用基于面向服务思想的Web架构解决这种具有动态、多变和复杂的疑难问题，于是伴随着利用面向服务思想设计的架构和软件集为一体的服务技术快速发展，运用面向服务思想的软件开发开始迅速地发展起来。在这种环境发展下，互联网上Web服务的规模开始迅速扩大，如WWW 2008的论文中统计发现：从2006年10月到2007年10月约一年的时间里，利用搜索引擎得到的WSDL服务的数量同比增长了286%。并且从2011年6月20日到2014年4月21日，世界著名服务注册网站ProgrammableWeb上的接收到的服务数目一直在稳步增长，并在不到三年的时间里服务数目由3261个增长为11222个，其增长幅度已经达到244%。此外，越来越多的企业也通过碎片化拆分和封装把原来的遗留软件系统转换为Web服务并部署到网络上。出于这样的原因，Web服务规模不断剧增，使大众用户难以准确、高效地发现Web服务。同时，互联网上发布的Web服务不论是利用XML进行描述的还是利用自然语言描述的服务，都存在着其描述特征词汇较少（例如PWeb中的服务描述特征词小于100个的占70%以上），并且在转换之后的特征词汇统计信息存在不足的问题。这种语义稀疏的现象使相似度计算无法有效的进行，此问题给次基于关键字的搜索方法带来了新的难题。

[0003] 因此，在互联网Web服务的规模化持续增长的环境下，如何根据现有的Web服

务描述语义稀疏的特点来进行准确、高效的Web服务发现，是现在服务计算领域中一个亟需解决的关键技术问题。

发明概述

技术问题

问题的解决方案

技术解决方案

[0004] 针对现有的Web服务描述语义稀疏的特点，为了能够实现准确、高效的Web服务发现，本发明提供了一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法。

[0005] 本发明采用以下的技术方案：

[0006] 一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法，包括以下步骤：

[0007] 步骤1：收集Web服务描述文档，对收集的Web服务描述文档进行解析，得到Web服务描述文档中的特征词汇，并对Web服务描述文档中的特征词汇进行预处理，得到一组原型单词的集合；

[0008] 步骤2：使用词嵌入训练模型Word2Vec训练步骤1中得到的原型单词的集合，得到原型单词的集合中每个单词的连续向量表示；

[0009] 使用Gaussian ATM模型对步骤1中得到的原型单词的集合进行训练，获得每个Web服务的层次结构；

[0010] 步骤3：使用步骤2训练得到的连续向量的集合丰富用户查询，得到扩展以后的用户查询；

[0011] 利用步骤2得到的服务层次结构，使用概率排序方法，得到针对扩展之后用户查询所对应的查询输出。

[0012] 优选地，步骤1中所述的Web服务描述文档来源于能够被访问的Web服务注册中心或Web服务门户网站。

[0013] 优选地，所述步骤1中的预处理包括去除停用词、取得单词词根和扩充缩写词。

[0014] 优选地，所述的步骤2，其具体实现包括以下子步骤：

[0015] 步骤2.1：使用Word2Vec模型训练原型单词的集合，得到每个单词的连续词嵌入表达；

[0016] 步骤2.2：基于步骤2.1中得到的词嵌入进行映射，遍历服务描述中的每一个单词，如果单词在词嵌入表达集合中，则直接使用词嵌入代替该单词；如果单词不在词嵌入表达集合中，该单词将被丢弃；

[0017] 经过多次迭代将服务描述中的单词转换成为词嵌入表达，从而将服务特征表示到连续词嵌入空间，为词嵌入主题模型Gaussian ATM构建输入；

[0018] 步骤2.3：以步骤2.2得到的服务描述的词嵌入表示为输入，基于Gaussian ATM，挖掘服务知识，构建服务-标签、标签-主题和主题-词嵌入分布，基于上述三个分布，建立服务-标签-主题-词嵌入层次模型，从而将服务组织到不同的主题。

[0019] 优选地，所述步骤2.3，具体包括以下子步骤：

[0020] 步骤2.3.1：在Gaussian ATM中，每个单词

w

对应的词嵌入

$$v(w) \in \mathbb{R}^M$$

，其中

M

表示词嵌入的长度，这样每个单词就被表示到一个

M

维的连续空间，而每个主题

z

就被表示成词嵌入

v

上的一个多元高斯分布，其中均值为

μ_z

，协方差为

$$\Sigma_x$$

，以服务的词嵌入表达作为输入，经过Gibbs抽样，算法运行收敛之后，通过该多元高斯分布，建立主题-词嵌入层次；

[0021] 步骤2.3.2: 每个服务文档可以当作标签的均匀分布，它们的分布是预先可观测的，直接利用均匀分布的计算公式可以获得服务-标签分布；

[0022] 步骤 2.3.3: 每个主题属于不同的标签，建立标签-主题层次的方式与步骤2.3.1类似，每个主题的先验分布为参数为

$$\alpha$$

的Dirichlet分布，以标签的词嵌入表达作为输入，经过Gibbs抽样之后得到收敛后的参数值，从而根据参数

$$\theta$$

获得标签-主题分布；

[0023] 步骤2.3.4: 步骤2.3.3之后，单词、词嵌入、标签、主题和服务文档存在如下关系：每个单词都有一个对应的词嵌入，每个词嵌入属于不同的主题，每个主题又隶属于不同的标签，每个标签又属于不同的服务，根据上述关系，以获得的服务-标签、标签-主题、主题-词嵌入分布作为输入，建立服务-标签-主题-词嵌入层次模型，从而将服务组织到不同的潜在主题。

[0024] 优选地，所述步骤3中，具体包括以下子步骤：

[0025] 步骤3.1: 根据查询特征表达，使用步骤2.1训练的连续词嵌入表达，遍历用户查询中的每一个单词，如果单词在词嵌入表达集合中，则直接使用词嵌入代替该单词；如果单词不在词嵌入表达集合中，该单词将被丢弃；将用户查询映射为连续表达的词嵌入，为基于词嵌入的查询语义扩充和基于词嵌入主题模型的服务发现提供支持；

[0026] 步骤3.2: 根据步骤3.1中获得用户查询的词嵌入表示作为输入，并设定一个相似度阈值，遍历输入中的每一个词嵌入，检索与该词嵌入的相似度超过预定阈值的词嵌入，将该词嵌入作为扩充词保留，重复该过程直至输入中所有的词嵌

入都被遍历一遍，最后返回添加了扩充词之后的用户查询向量；

[0027] 步骤3.3: 词嵌入表达的用户查询将基于一种概率表达映射为同服务的相似程度，根据服务-主题-词嵌入层次结构可以获得词嵌入同主题的相似度：

$$P(e|z)$$

，每个查询

$$Q$$

同主题的相似度能够表示为公式(1)：

[0028]

$$P(Q|z_f) = \prod_{e_i \in Q} P(e_i|z_f) \quad (1)$$

[0029] 其中

$$z_f$$

为主题，

$$e_i$$

为查询中包含的词嵌入；

[0030] 同样根据服务-主题-词嵌入层次结构，可以获得主题同服务的相似度：

$$P(z|t)$$

，每个查询

$$Q$$

同服务

$$t_i$$

的相似程度如公式(2)所示：

[0031]

$$P(Q|t_i) = \prod_{e_x \in Q} P(e_x|t_i) = \prod_{e_x \in Q} \sum_{z_f=1}^Z P(e_x|z_f) P(z_f|t_i)$$

(2)

[0032] 其中

 Q

为查询,

 t_i

为标签,

 z_f

为潜在特征,

 e_x

为查询包含的词嵌入,

 Z

为主题数目; 的取值表示查询和服务的相似程度, 将被用来为用户提供服务匹配和排序的依据;

[0033] 最后, 计算每个查询

 Q

同服务

 s_j

的相似程度

 $P(Q|s_j)$

, 该概率的大小将被用来为用户提供服务匹配和排序的依据;

[0034]

$$P(Q|s_j) = \prod_{e_x \in Q} P(e_x|s_j) = \prod_{e_x \in Q} \sum_{f=1}^Z \sum_{i=1}^T P(e_x|z_f) P(z_f|t_i) P(t_i|s_j)$$

(3)

[0035] 其中

Q

为查询,

 s_j

为服务,

 t_i

为标签,

 z_j

为潜在特征,

 e_q

为查询包含的词嵌入,

 Z

为主题数目,

 T

为标签的数量;

 $P(Q|s_j)$

的取值表示查询和服务的相似程度, 将被用来为用户提供服务匹配和排序的依据。

发明的有益效果

有益效果

[0036] 本发明具有的有益效果是:

[0037] (1) 可以实现语义稀疏Web服务的组织和检索, 具有较好的通用性;

[0038] (2) 基于“词袋”模型的服务描述特征表达方法存在“维数过多”的问题, 本方法通过引入词嵌入模型来将服务描述的表达映射到固定长度的词嵌入空间从而确保服务描述特征的维度, 本方法在保证维数确定的情况下还能够利用词嵌入

的语义相似特性增强服务描述的语义；

[0039] (3) 因为本方法采用的是无监督的学习方法，所以参数推导不需要过多的经验值，提高了方法的准确性；

[0040] (4) 学习方法建立的服务层次可以用于Web服务层次组织，从而支撑语义稀疏Web服务发现，具有广泛的适用性。

对附图的简要说明

附图说明

[0041] 图1为基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法的流程示意图。

[0042] 图2为Gaussian ATM生成的Web服务层次示意图。

[0043] 图3为运行步骤2之后得到的Web服务层次。

发明实施例

本发明的实施方式

[0044] 下面结合附图和具体实施例对本发明的具体实施方式做进一步说明：

[0045] 结合图1至图3，一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法，包括以下步骤：

[0046] 步骤1：收集Web服务描述文档，对收集的Web服务描述文档进行解析，得到Web服务描述文档中的特征词汇，并对Web服务描述文档中的特征词汇进行预处理，得到一组原型单词的集合。

[0047] Web服务描述文档来源于能够被访问的Web服务注册中心或Web服务门户网站。

[0048] 预处理包括去除停用词、取得单词词根和扩充缩写词等。

[0049] 步骤2：使用词嵌入训练模型Word2Vec训练步骤1中得到的原型单词的集合，得到原型单词的集合中每个单词的连续向量表示；

[0050] 使用Gaussian ATM (Author-Topic-Model) 模型对步骤1中得到的原型单词的集合进行训练，获得每个Web服务的层次结构。

[0051] 其具体实现包括以下子步骤：

[0052] 步骤2.1：使用Word2Vec模型训练原型单词的集合，得到每个单词的连续词嵌

入表达；

[0053] 步骤2.2：基于步骤2.1中得到的词嵌入进行映射，遍历服务描述中的每一个单词，如果单词在词嵌入表达集合中，则直接使用词嵌入代替该单词；如果单词不在词嵌入表达集合中，该单词将被丢弃；

[0054] 经过多次迭代将服务描述中的单词转换成为词嵌入表达，从而将服务特征表示到连续词嵌入空间，为词嵌入主题模型Gaussian ATM构建输入；

[0055] 步骤2.3：以步骤2.2得到的服务描述的词嵌入表示为输入，基于Gaussian ATM，挖掘服务知识，构建服务-标签、标签-主题和主题-词嵌入分布，基于上述三个分布，建立服务-标签-主题-词嵌入层次模型，从而将服务组织到不同的主题。

[0056] 步骤2.3，具体包括以下子步骤：

[0057] 步骤2.3.1：在Gaussian ATM中，每个单词

w

对应的词嵌入

$$v(w) \in \mathbb{R}^M$$

，其中

M

表示词嵌入的长度，这样每个单词就被表示到一个

M

维的连续空间，而每个主题

z

就被表示成词嵌入

v

上的一个多元高斯分布，其中均值为

μ_z

，协方差为

$$\sum_z$$

，以服务的词嵌入表达作为输入，经过Gibbs抽样，算法运行收敛之后，通过该多元高斯分布，建立主题-词嵌入层次；

[0058] 步骤2.3.2: 每个服务文档可以当作标签的均匀分布，它们的分布是预先可观测的，直接利用均匀分布的计算公式可以获得服务-标签分布；

[0059] 步骤 2.3.3: 每个主题属于不同的标签，建立标签-主题层次的方式与步骤2.3.1类似，每个主题的先验分布为参数为

$$\alpha$$

的Dirichlet分布，以标签的词嵌入表达作为输入，经过Gibbs抽样之后得到收敛后的参数值，从而根据参数

$$\theta$$

获得标签-主题分布；

[0060] 步骤2.3.4: 步骤2.3.3之后，单词、词嵌入、标签、主题和服务文档存在如下关系：每个单词都有一个对应的词嵌入，每个词嵌入属于不同的主题，每个主题又隶属于不同的标签，每个标签又属于不同的服务，根据上述关系，以获得的 服务-标签、标签-主题、主题-词嵌入分布作为输入，建立服务-标签-主题-词嵌入层次模型，从而将服务组织到不同的潜在主题。

[0061] 如图2所示，图2是本发明的执行Gaussian ATM算法之后的生成的服务层次模型示意图。

[0062] 在图2中，每个单词都有一个对应的词嵌入，每个词嵌入属于不同的主题，每个主题又隶属于不同的标签，每个标签又属于不同的服务。基于这个三个层次结构可以为用户提供灵活的服务查询体验。例如可以基于主题组织服务，单词分布类似的服务将划分到同一个主题从而将服务聚类。通过在类簇上查询服务减小搜索空间，提高查询效率。而在本发明中，我们利用该层次结构建立一种概率服务发现方法。

[0063] 步骤3: 使用步骤2训练得到的连续向量的集合丰富用户查询，得到扩展以利用步骤2得到的服务层次结构，使用概率排序方法，得到针对扩展之后用户查询所

对应的查询输出。

[0064] 具体包括以下子步骤:

[0065] 步骤3.1: 根据查询特征表达, 使用步骤2.1训练的连续词嵌入表达, 遍历用户查询中的每一个单词, 如果单词在词嵌入表达集合中, 则直接使用词嵌入代替该单词; 如果单词不在词嵌入表达集合中, 该单词将被丢弃; 将用户查询映射为连续表达的词嵌入, 为基于词嵌入的查询语义扩充和基于词嵌入主题模型的服务发现提供支持;

[0066] 步骤3.2: 根据步骤3.1中获得用户查询的词嵌入表示作为输入, 并设定一个相似度阈值, 遍历输入中的每一个词嵌入, 检索与该词嵌入的相似度超过预定阈值的词嵌入, 将该词嵌入作为扩充词保留, 重复该过程直至输入中所有的词嵌入都被遍历一遍, 最后返回添加了扩充词之后的用户查询向量;

[0067] 步骤3.3: 词嵌入表达的用户查询将基于一种概率表达映射为同服务的相似程度, 根据服务-主题-词嵌入层次结构可以获得词嵌入同主题的相似度:

$$P(e|z)$$

, 每个查询

Q

同主题的相似度能够表示为公式(1):

[0068]

$$P(Q|z_f) = \prod_{e_i \in Q} P(e_i|z_f)$$

(1)

[0069] 其中

z_f

为主题,

e_i

为查询中包含的词嵌入;

[0070] 同样根据服务-主题-词嵌入层次结构, 可以获得主题同服务的相似度:

$$P(z|t)$$

, 每个查询

$$Q$$

同服务

$$t_i$$

的相似程度如公式(2)所示:

[0071]

$$P(Q|t_i) = \prod_{e_x \in Q} P(e_x | t_i) = \prod_{e_x \in Q} \sum_{z_f=1}^Z P(e_x | z_f) P(z_f | t_i)$$

(2)

[0072] 其中

$$Q$$

为查询,

$$t_i$$

为标签,

$$z_f$$

为潜在特征,

$$e_x$$

为查询包含的词嵌入,

$$Z$$

为主题数目; 的取值表示查询和服务的相似程度, 将被用来为用户提供服务匹配和排序的依据;

[0073] 最后, 计算每个查询

$$Q$$

同服务

s_j

的相似程度

$P(Q|s_j)$

，该概率的大小将被用来为用户提供服务匹配和排序的依据；

[0074]

$$P(Q|s_j) = \prod_{e_x \in Q} P(e_x | s_j) = \prod_{e_x \in Q} \sum_{z_f=1}^Z \sum_{t_i=1}^T P(e_x | z_f) P(z_f | t_i) P(t_i | s_j)$$

(3)

[0075] 其中

Q

为查询，

s_j

为服务，

t_i

为标签，

z_f

为潜在特征，

e_x

为查询包含的词嵌入，

Z

为主题的数量，

T

为标签的数量；

$$P(Q|s_j)$$

的取值表示查询和服务的相似程度，将被用来为用户提供服务匹配和排序的依据。

[0076] 下面是应用本发明的方法进行语义稀疏服务发现的具体实施例：以ProgrammableWeb (PWeb, <http://www.programmableweb.com/>) 网站上的Web服务进行基于Web服务描述的语义稀疏服务发现为实施例，详细描述本发明的实施过程。

[0077] PWeb网站是目前互联网上可以公开访问的著名mashup和Web服务注册中心，截至2016年5月1日，注册的Web服务已超过15000个，包括遵循SOAP、REST、XML-RPC的各类Web API，并且提供了API服务的一些注册信息，比如API描述、名字、标签等。对于采用SOAP协议的API服务，还提供了描述该Web服务的WSDL文档。

[0078] 首先，执行步骤1，使用PWeb网站提供的开放API从该网站收集Weather、Government等领域的服务自然语义描述，并对其进行预处理。

[0079] 然后，对收集的Web服务描述文档进行解析，具体包括：对收集的每个Web服务描述文档进行解析，抽取该文档中的所有核心特征词汇，并将词汇做进一步处理，包括词干还原、缩写词扩充等。

[0080] 接着，执行步骤2，使用word2vec分布式表示方法的训练特征单词集合，建立每个单词的连续向量表示。在调用word2vec之后，系统将得到表1所示的结果。每个单词都获得一个连续向量的表达，向量的长度可以选择，本例中选择向量长度为50。

[0081] 表1 使用word2vec方法之后，单词的连续向量表达

[]

[表1]

单词	连续向量
width	-0.41559 -0.148131 -0.0797283 0.0545502 -0.21033 -0.117021 0.298572 -0.0717266 -0.00524523 0.0592786 -0.169959 -0.128289 -0.186867 0.0266605 0.253384 0.0223827 -0.366654 0.216775 -0.0591345 0.0317087 -0.149114 -0.0559157 -0.123391 -0.0138674 -0.086514 0.0592855 0.1126421 -0.212031 -0.0543314 0.0720561 0.14411 -0.0839934 0.0073932 0.0586023 0.0272626 0.0589208 0.0521655 -0.0755701 0.225914 -0.0895848 0.244108 0.121313 0.0116312 0.161205 -0.0843894 0.0510214 -0.111115 0.108711 -0.0516607
ship	0.211417 -0.234613 -0.0241331 -0.0521823 -0.0147823 -0.205011 -0.0560595 0.315565 -0.129355 0.01111573 0.0545653 -0.141668 -0.132769 -0.112211 0.0848746 0.205129 0.0779571 -0.369922 0.211233 0.00521666 0.0115437 -0.01118886 -0.208547 -0.110523 0.156774 -0.121732 0.0251678 0.0122169 -0.241777 -0.00805405 0.135456 0.118953 -0.0566561 -0.159758 0.0139899 0.0899661 0.207916 -0.00802972 -0.1157902 0.286131 -0.0556772 0.211616 0.110553 0.1151582 0.0771187 0.0244805 0.0835618 -0.114896 0.145608 -0.0822288

[0082]

[0083] 训练得到每个单词的连续表示之后，使用Gaussian ATM模型建立Web服务的层次。建立的服务—主题层次如表2所示。

[0084] 表2 使用Gaussian ATM方法，建立的服务—标签层次

[]

[表2]

文档序号	标签的概率
1812	0.03 0.12 0.12 0.013 0.12 0.12 0.51 0.12 0.12 0.81...
1277	0.15 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.89...
...	...

[0085]

[0086] 表3 使用Gaussian ATM方法，建立的标签－主题层次

[] [表3]

标签序号	包含主题的概率
812	0.0012 0.0012 0.0012 0.2113 0.0012 0.0012 0.042 0.0012 0.0012 0.81
277	0.0044 0.0044 0.0044 0.0044 0.0044 0.0044 0.0044 0.0044 0.0044 0.89
...	...

[0087]

[0088] 建立的主题－单词层次如表3所示。

[0089] [0092] 表4 使用Gaussian ATM方法，建立的主题－词嵌入层次

[] [表4]

主题序号	包含单词的概率
1	0:0.051 1:0.06 2:0.051 3:0.06 4:0.051 5:0.031 6:0.031 7:0.089 8:0.057 9:0.031 10:0.063...
2	0:0.021 1:0.051 2:0.044 3:0.056 4:0.046 5:0.031 6:0.016 7:0.069 8:0.06 9:0.034 10:0.05...
...	...

[0090] 然后执行步骤3，利用连续向量的相似特性，扩展用户查询。例如，design扩展为design + plan。利用步骤2得到的服务层次结构，使用概率排序方法，得到针对扩展之后用户查询所对应的查询输出，最终返回的结果如表5所示。

[0091] 表5 查询与系统输出结果

[0092]

查询关键词	包含top-10单词及对应概率
weather	OpenWeatherMap 5 Day Weather Forecast: 0.95
forecast	Metwit Weather and Forecast:0.90
	AccuWeather Forecast: 0.88
	MSN Forecast: 0.88
...	...

[0093] 请见图3，运行步骤2之后得到的Web服务层次如图3所示。

[0094] 当然，上述说明并非是对本发明的限制，本发明也并不仅限于上述举例，本技术领域的技术人员在本发明的实质范围内所做出的变化、改型、添加或替换，也应属于本发明的保护范围。

权利要求书

- [权利要求 1] 一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法，其特征在于，包括以下步骤：
- 步骤1：收集Web服务描述文档，对收集的Web服务描述文档进行解析，得到Web服务描述文档中的特征词汇，并对Web服务描述文档中的特征词汇进行预处理，得到一组原型单词的集合；
- 步骤2：使用词嵌入训练模型Word2Vec训练步骤1中得到的原型单词的集合，得到原型单词的集合中每个单词的连续向量表示；
- 使用Gaussian ATM模型对步骤1中得到的原型单词的集合进行训练，获得每个Web服务的层次结构；
- 步骤3：使用步骤2训练得到的连续向量的集合丰富用户查询，得到扩展以后的用户查询；
- 利用步骤2得到的服务层次结构，使用概率排序方法，得到针对扩展之后用户查询所对应的查询输出。
- [权利要求 2] 根据权利要求1所述的一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法，其特征在于，步骤1中所述的Web服务描述文档来源于能够被访问的Web服务注册中心或Web服务门户网站。
- [权利要求 3] 根据权利要求1所述的一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法，其特征在于，所述步骤1中的预处理包括去除停用词、取得单词词根和扩充缩写词。
- [权利要求 4] 根据权利要求1所述的一种基于Gaussian ATM和词嵌入的语义稀疏Web服务发现方法，其特征在于，所述的步骤2，其具体实现包括以下子步骤：
- 步骤2.1：使用Word2Vec模型训练原型单词的集合，得到每个单词的连续词嵌入表达；
- 步骤2.2：基于步骤2.1中得到的词嵌入进行映射，遍历服务描述中的每一个单词，如果单词在词嵌入表达集合中，则直接使用词嵌入代替

该单词；如果单词不在词嵌入表达集合中，该单词将被丢弃；

经过多次迭代将服务描述中的单词转换成为词嵌入表达，从而将服务特征表示到连续词嵌入空间，为词嵌入主题模型Gaussian ATM构建输入；

步骤2.3：以步骤2.2得到的服务描述的词嵌入表示为输入，基于Gaussian ATM，挖掘服务知识，构建服务-标签、标签-主题和主题-词嵌入分布，基于上述三个分布，建立服务-标签-主题-词嵌入层次模型，从而将服务组织到不同的主题。

[权利要求 5]

根据权利要求4所述的一种基于Gaussian

ATM和词嵌入的语义稀疏Web服务发现方法，其特征在于，所述步骤2.3，具体包括以下子步骤：

步骤2.3.1：在Gaussian ATM中，每个单词

w

对应的词嵌入

$v(w) \in \mathbb{R}^M$

，其中

M

表示词嵌入的长度，这样每个单词就被表示到一个

M

维的连续空间，而每个主题

z

就被表示成词嵌入

v

上的一个多元高斯分布，其中均值为

μ_z

，协方差为

$$\sum_z$$

，以服务的词嵌入表达作为输入，经过Gibbs抽样，算法运行收敛之后，通过该多元高斯分布，建立主题-词嵌入层次；

步骤2.3.2：每个服务文档可以当作标签的均匀分布，它们的分布是预先可观测的，直接利用均匀分布的计算公式可以获得服务-标签分布；

步骤 2.3.3：每个主题属于不同的标签，建立标签-主题层次的方式与步骤2.3.1类似，每个主题的先验分布为参数为

$$\alpha$$

的Dirichlet分布，以标签的词嵌入表达作为输入，经过Gibbs抽样之后得到收敛后的参数值，从而根据参数

$$\theta$$

获得标签-主题分布；

步骤2.3.4：步骤2.3.3之后，单词、词嵌入、标签、主题和服务文档存在如下关系：每个单词都有一个对应的词嵌入，每个词嵌入属于不同的主题，每个主题又隶属于不同的标签，每个标签又属于不同的服务，根据上述关系，以获得的服务-标签、标签-主题、主题-词嵌入分布作为输入，建立服务-标签-主题-词嵌入层次模型，从而将服务组织到不同的潜在主题。

[权利要求 6]

根据权利要求4所述的一种基于Gaussian

ATM和词嵌入的语义稀疏Web服务发现方法，其特征在于，所述步骤3中，具体包括以下子步骤：

步骤3.1：根据查询特征表达，使用步骤2.1训练的连续词嵌入表达，遍历用户查询中的每一个单词，如果单词在词嵌入表达集合中，则直接使用词嵌入代替该单词；如果单词不在词嵌入表达集合中，该单词将被丢弃；将用户查询映射为连续表达的词嵌入，为基于词嵌入的查询语义扩充和基于词嵌入主题模型的服务发现提供支持；

步骤3.2: 根据步骤3.1中获得用户查询的词嵌入表示作为输入, 并设定一个相似度阈值, 遍历输入中的每一个词嵌入, 检索与该词嵌入的相似度超过预定阈值的词嵌入, 将该词嵌入作为扩充词保留, 重复该过程直至输入中所有的词嵌入都被遍历一遍, 最后返回添加了扩充词之后的用户查询向量;

步骤3.3: 词嵌入表达的用户查询将基于一种概率表达映射为同服务的相似程度, 根据服务-主题-词嵌入层次结构可以获得词嵌入同主题的相似度:

$$P(e|z)$$

, 每个查询

$$Q$$

同主题的相似度能够表示为公式(1):

$$P(Q|z_f) = \prod_{e \in Q} P(e_i|z_f)$$

(1)

其中

$$z_f$$

为主题,

$$e_i$$

为查询中包含的词嵌入;

同样根据服务-主题-词嵌入层次结构, 可以获得主题同服务的相似度:

$$P(z|t)$$

, 每个查询

$$Q$$

同服务

t_i

的相似程度如公式(2)所示:

$$P(Q|t_i) = \prod_{e_x \in Q} P(e_x | t_i) = \prod_{e_x \in Q} \sum_{z_f=1}^Z P(e_x | z_f) P(z_f | t_i) \quad (2)$$

其中

Q

为查询,

t_i

为标签,

z_f

为潜在特征,

e_x

为查询包含的词嵌入,

Z

为主题数目; t_i 的取值表示查询和服务的相似程度, 将被用来为用户提供服务匹配和排序的依据;

最后, 计算每个查询

Q

同服务

s_j

的相似程度

$$P(Q|s_j)$$

，该概率的大小将被用来为用户提供服务匹配和排序的依据；

$$P(Q|s_j) = \prod_{e_k \in Q} P(e_k | s_j) = \prod_{e_k \in Q} \sum_{f=1}^Z \sum_{t=1}^T P(e_k | z_f) P(z_f | t) P(t | s_j)$$

(3)

其中

$$Q$$

为查询，

$$s_j$$

为服务，

$$t_i$$

为标签，

$$z_f$$

为潜在特征，

$$e_k$$

为查询包含的词嵌入，

$$Z$$

为主题数目，

$$T$$

为标签的数量；

$$P(Q|s_j)$$

的取值表示查询和服务的相似程度，将被用来为用户提供服务匹配和排序的依据。

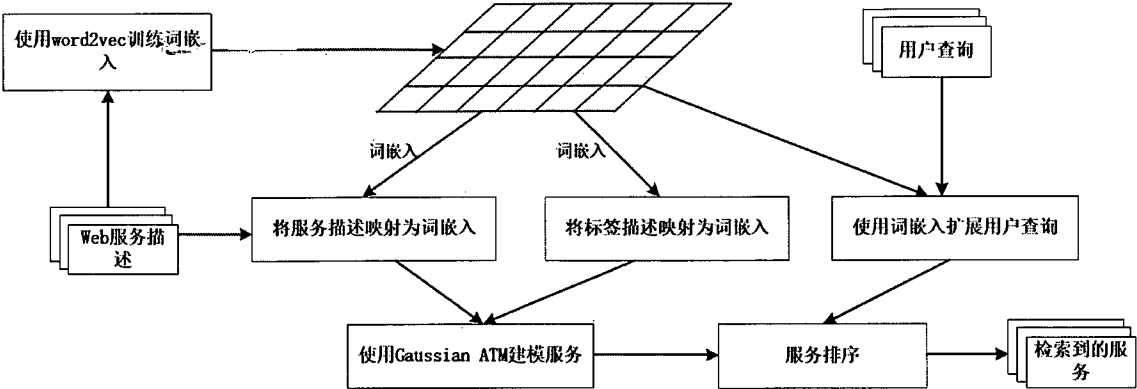


图 1

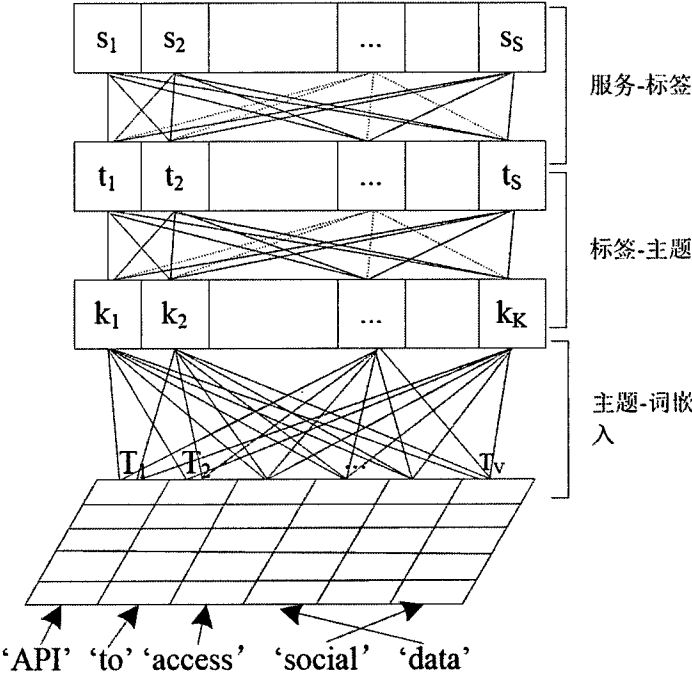


图 2

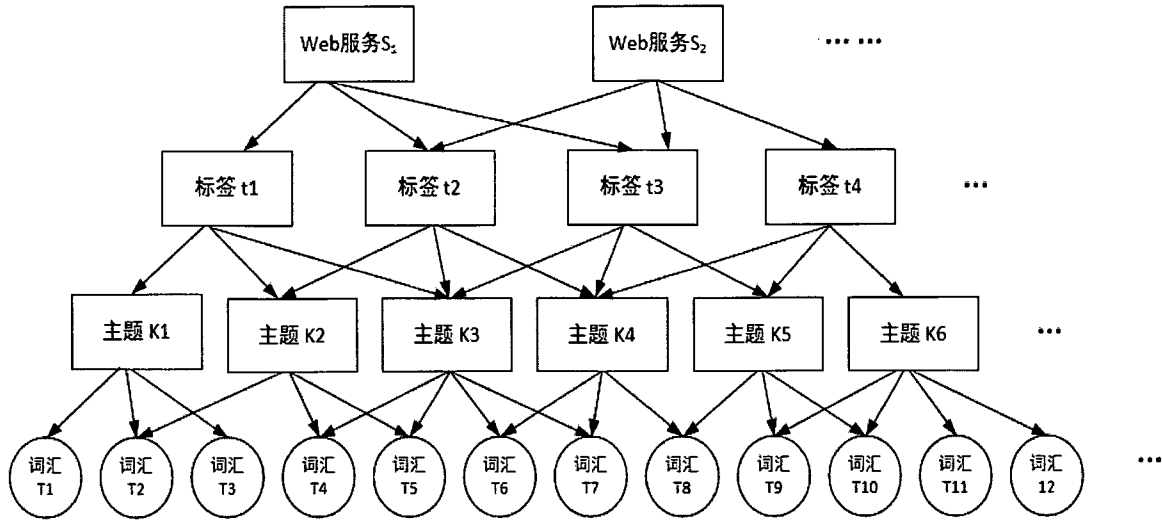


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/079967

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, VEN, CNTXT, USTXT, WOTXT, EPTXT, CNKI: 线性判别分析, 作者主题模型, 语义, Web服务, 发现; LDA, ATM, semantic, discover+, service, Web

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106599086 A (SHANDONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 26 April 2017 (2017-04-26) claims 1-6	1-6
PX	CN 108712466 A (SHANDONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 26 October 2018 (2018-10-26) claims 1-6	1-6
A	CN 106156272 A (BEIJING UNIVERSITY OF TECHNOLOGY) 23 November 2016 (2016-11-23) entire document	1-6
A	US 2011270870 A1 (IBM) 03 November 2011 (2011-11-03) entire document	1-6

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

23 May 2019

Date of mailing of the international search report

04 June 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/079967

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106599086	A	26 April 2017	None			
CN	108712466	A	26 October 2018	None			
CN	106156272	A	23 November 2016	None			
US	2011270870	A1	03 November 2011	US	2012221598	A1	30 August 2012
				US	8667011	B2	04 March 2014
				US	8275806	B2	25 September 2012

国际检索报告

国际申请号

PCT/CN2019/079967

A. 主题的分类

H04L 29/08(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04L; G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNABS, VEN, CNTXT, USTXT, WOTXT, EPTXT, CNKI:线性判别分析, 作者主题模型, 语义, Web服务, 发现; LDA, ATM, semantic, discover+, service, Web

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 106599086 A (山东科技大学) 2017年 4月 26日 (2017 - 04 - 26) 权利要求1-6	1-6
PX	CN 108712466 A (山东科技大学) 2018年 10月 26日 (2018 - 10 - 26) 权利要求1-6	1-6
A	CN 106156272 A (北京工业大学) 2016年 11月 23日 (2016 - 11 - 23) 全文	1-6
A	US 2011270870 A1 (IBM) 2011年 11月 3日 (2011 - 11 - 03) 全文	1-6

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 5月 23日

国际检索报告邮寄日期

2019年 6月 4日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

受权官员

卞晓飞

传真号 (86-10)62019451

电话号码 86-10-62411330

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/079967

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	106599086	A	2017年 4月 26日	无			
CN	108712466	A	2018年 10月 26日	无			
CN	106156272	A	2016年 11月 23日	无			
US	2011270870	A1	2011年 11月 3日	US	2012221598	A1	2012年 8月 30日
				US	8667011	B2	2014年 3月 4日
				US	8275806	B2	2012年 9月 25日

表 PCT/ISA/210 (同族专利附件) (2015年1月)