

基于BTM的物联网服务发现方法

王舒漫, 李爱萍*, 段利国, 付佳, 陈永乐

(太原理工大学 信息与计算机学院, 太原 030024)

(* 通信作者电子邮箱 tyutli@163.com)

摘要:针对物联网(IoT)服务描述文本篇幅较短、特征稀疏,直接采用传统的主题模型对IoT服务建模得到的聚类效果不佳,从而导致无法发现最佳服务的问题,提出了一种基于BTM的IoT服务发现方法。该方法首先利用BTM挖掘现有IoT服务的隐含主题,并通过全局主题分布和主题-词分布计算推理得到服务文档-主题概率分布;其次利用K-means算法对服务进行聚类,并返回服务请求的最佳匹配结果。实验结果分析表明,该方法能够有效提高IoT服务的聚类效果,从而得到匹配的最佳服务。与现有的HDP(Hierarchical Dirichlet Process)、基于K-means的隐狄利克雷分配(LDA-K)等方法相比,该方法进行最佳服务发现的准确度(Precision)和归一化折损累积增益(NDCG)均有一定幅度的提高。

关键词:物联网服务;BTM;短文本;主题建模;服务发现

中图分类号:TP391 **文献标志码:**A

Service discovery method for Internet of Things based on Biterm topic model

WANG Shuman, LI Aiping*, DUAN Ligu, FU Jia, CHEN Yongle

(College of Information and Computer, Taiyuan University of Technology, Taiyuan Shanxi 030024, China)

Abstract: Service description texts for Internet of Things (IoT) are short in length and sparse in text features, and direct modeling the IoT service by using traditional topic model has poor clustering effect, so that the best service cannot be discovered. To solve this problem, an IoT service discovery method based on Biterm Topic Model (BTM) was proposed. Firstly, BTM was employed to mine the latent topic of the existing IoT services, and the service document-topic probability distribution was calculated and deduced through global topic distribution and theme-word distribution. Then, K-means algorithm was used to cluster the services and return the best matching results of service requests. Experimental results show that the proposed method can improve the clustering effect of services for IoT and thus obtain the matched best service. Compared with the methods of HDP (Hierarchical Dirichlet Process) and LDA-K (Latent Dirichlet Allocation based on K-means), the proposed method achieves better performance in terms of Precision and Normalized Discounted Cumulative Gain (NDCG) for best service discovery.

Key words: service for Internet of Things (IoT); Biterm Topic Model (BTM); short text; topic modeling; service discovery

0 引言

物联网(Internet of Things, IoT)作为我国战略性新兴产业的一个重要组成部分,正进入深化应用的新阶段^[1]。随着5G技术的推广和物联网设备的平价化,预计到2020年会有500亿设备接入物联网^[1],同时,也导致用户从数量巨大的物联网设备中精确地定位满足其特定业务需求的设备变得越来越困难和耗时,如何快速高效地辅助用户发现满足需求的设备成为了目前急需解决的问题。

基于面向服务架构(Service-Oriented Architecture, SOA)的物联网使物联网的功能服务化^[2],服务聚类是支持服务快速发现的有效辅助方法,聚类算法根据某一相似特性快速聚

合相关服务,将数据划分成不同的集合,能提高服务发现的效率。在物联网服务发现中,Liu等^[3]通过提取物联网服务的主题签名,利用主题建模方法进行聚类来确定相似服务集。Jiang等^[4]使用自然语言处理技术提取服务描述文本中的目标特征,计算两个服务目标特征之间的语义相似度,并采用K-means算法对服务进行聚类。

上述研究为物联网服务的发现提供了有益的方法参考。然而,物联网服务描述文档属于短文本,特征稀疏,信息量少,采用已有的服务发现方法进行聚类时会出现“类型无法识别、类型丢失”等现象,达不到理想效果。针对物联网服务的上述特点,本文提出了一种基于词对主题模型BTM(Biterm Topic Model)的物联网服务发现方法,根据服务的隐含主题进行

收稿日期:2019-08-12;修回日期:2019-09-29;录用日期:2019-10-24。

基金项目:国家重点研发计划“网络空间安全”专项子课题资助项目(2018YFB0803402)。

作者简介:王舒漫(1994—),女,山西长治人,硕士研究生,主要研究方向:服务计算、信息处理;李爱萍(1974—),女,山西文水人,教授,博士,CCF高级会员,主要研究方向:物联网数据处理、无线传感网络;段利国(1970—),男,山西繁峙人,副教授,博士,CCF高级会员,主要研究方向:中文信息处理、知识图谱;付佳(1995—),女,内蒙古通辽人,硕士研究生,主要研究方向:服务计算、信息处理;陈永乐(1983—),男,山东潍坊人,副教授,博士,CCF会员,主要研究方向:无线传感网络、物联网安全。

聚类。

1 相关工作

物联网的概念最早可追溯到 20 世纪 90 年代,国际电信联盟在 2005 年的 ITU 互联网报告中正式给出“物联网”的概念:按照既定的协议,通过互联网实现物与物之间信息交换和通信,进而实现智能化识别、定位、跟踪、监控和管理的一种网络。

1.1 物联网场景中的“物”的标识

物联网场景中的“物”主要指的是现实环境中存在的感知识别、信息交互和智能控制的任何设备和资源。物联网是开放式的、动态变化式的、高分布式的,允许设备随时在不同地点以多种方式接入物联网中;然而对设备信息没有统一格式的描述,从而导致接入物联网的设备呈现多种形式的信息表达,设备间信息的交互和共享受限^[5]。对物联网设备信息进行统一的资源描述,是面向物联网设备服务发现的基础。近年来,一些研究人员通过以下方法来解决目前物联网领域设备的统一描述问题。王书龙等^[6]提出了基于本体的物联网设备资源描述模型,根据设备特点将其划分为 5 类资源进行综合描述,包括属性、控制、状态、历史信息和隐私。Santos 等^[7]提出了一种针对嵌入式资源的描述语言,该描述语言从功能、需求和限制三方面来描述设备资源。结合以上研究,本文根据异构物联网设备的共有特征,结合物联网服务发现的需求,选取功能、接口、工作状态、工作环境四部分来构建物联网服务描述模型,如图 1 所示。其中,功能用于描述用户对服务功能的要求;接口用于描述用户对于相关的设备服务接入方式的要求,包括接口信息、通信方式等;工作状态用于描述用户对于服务相关联的设备的当前运行状态的要求;工作环境用于描述用户对于服务相关联的设备的正常工作环境的要求,包括工作温度、湿度、地理位置等。

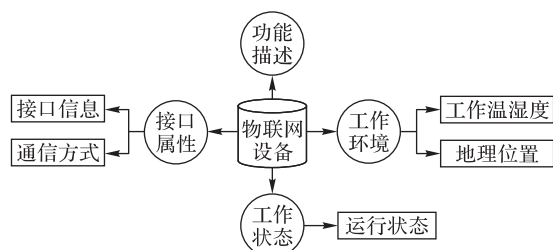


图1 物联网服务描述文本

Fig. 1 Service description text for IoT

1.2 现有的物联网服务发现方法

随着技术发展,物联网服务数量还在逐步呈不断上升的趋势,高效快速地发现最佳的服务成为亟待解决的问题。

通过挖掘服务的潜在语义信息对服务进行聚类,尤其是借助于主题模型进行服务发现,能有效地提高服务聚类的精度和速度,减少资源消耗。在主题模型中,认为一个服务描述文本包含多个隐含主题。Teixeira 等^[8]提出采用概率发现的方法来进行服务发现,提高了服务发现的速度,并降低了服务资源的消耗。Casser 等^[9]提出了一种混合语义的服务匹配方法,采用概率主题模型来学习服务的隐含主题,并通过主题相似度进行服务聚类。同时,由于物联网设备故障或损坏、设备的移动性等因素,使得物联网服务经常出现消失或重现等动态变化特点,使用主题模型对物联网服务进行主题建模成为解决服务动态变化的有效方法之一。

然而,物联网服务描述文档属于短文本,信息量少,缺乏足够的词频共现,直接使用传统的主题模型或其扩展模型对短文本进行建模,会导致严重的数据稀疏问题。因此,对语义信息稀少的短文本进行合理的建模,是实现高质量服务发现的关键。

近年来,现有的研究中有部分尝试将短文本进行扩充来进行主题模型的训练。魏强等^[10]利用英文 Wikipedia 对短文本的服务描述进行语义扩充,将短文本转换为长文本进行建模来构建高质量的主题模型。肖巧翔等^[11]通过使用 word2vec 扩充服务描述语义,然后基于隐狄利克雷分配 (Latent Dirichlet Allocation, LDA) 模型对服务进行聚类。还有一些研究考虑利用外部知识库对短文本进行特征扩充,这也是一种较常见的方法,但恰当的外部数据集不容易找到,而且使用一个外部知识库会使建模精确度下降,同时,这些方法没有普遍性。2013 年, Yan 等^[12]在 LDA 和一元混合模型的基础上提出了基于词对信息的词对主题模型 BTM,在不进行文本扩充的情况下能有效地克服短文本数据稀疏问题,同时考虑了词之间的语义联系。因此,本文使用 BTM 对物联网服务的隐含主题进行挖掘,再通过主题进行服务聚类。

2 BTM

传统的主题模型模拟文本的生成过程,再通过参数估计得到隐含主题。与传统的主题模型相比, BTM 通过将文档转换为词对,直接对整个语料库的词对 biterm (即共现的无序词对模式) 进行建模来学习短文本中的主题。BTM 的图模型表示如图 2 所示。其中: φ_z 表示主题 z 下的词对概率分布, θ 表示语料库中全局主题概率分布, 多项式分布的参数 φ_z 和 θ 分别用于生成主题和词对; T 代表主题个数; $|B|$ 代表语料库 biterm 集合 B 中词对的总数; w_i, w_j 和 z 分别代表词对 b 的两个词及其主题; α 和 β 是 Dirichlet 先验分布的参数。

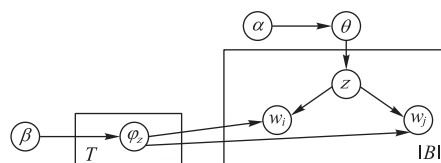


图2 BTM的图模型表示

Fig. 2 Graphical model representation of BTM

2.1 生成过程

在 BTM 中,语料库由多个主题组成,每个 biterm 独立地从特定主题中抽取,具体生成过程如下:

- 1) 生成各个主题 z 下的主题-词分布 $\varphi_{w|z} \sim \text{Dir}(\beta)$ 。
- 2) 生成整个语料库的全局的主题分布 $\theta_z \sim \text{Dir}(\alpha)$ 。
- 3) 生成 biterm 集合 B 中的每个词对 $b = (w_i, w_j)$:
 - a) 从全局的 θ 中抽取一个主题 $z \sim \text{Mult}(\theta)$;
 - b) 从主题 z 中抽取两个词: $w_i, w_j \sim \text{Mult}(\varphi_{w|z})$ 。

按照以上的生成过程,词对 $b = (w_i, w_j)$ 的联合概率可以表示为:

$$p(b) = \sum_z p(z) p(w_i, w_j | z) = \sum_z \theta_z \varphi_{i|z} \varphi_{j|z} \quad (1)$$

因此,产生 BTM 语料库的概率为:

$$p(B) = \prod_{(i,j)} \sum_z \theta_z \varphi_{i|z} \varphi_{j|z} \quad (2)$$

2.2 参数估计

在 BTM 中,最重要的两组参数是全局主题分布 θ_z 和主

题-词分布 $\varphi_{w|z}$, 参数估计可以看作是生成过程的逆过程: 在已知生成结果(即已知语料库)的情况下, 通过参数估计^[12]得到参数值。与LDA类似, BTM不能对参数进行精确推断, 因此采用Gibbs抽样^[13]进行近似推断。Gibbs抽样是一种简单而广泛适用的马尔可夫蒙特卡洛(Markov Chain Monte Carlo, MCMC)算法, 更渐进地接近正确的分布。执行Gibbs抽样, 首先选择MCMC的初始状态, 即把词对、词随机分配到各主题下, 然后计算每个词对 $b = (w_i, w_j)$ 的条件概率 $p(z|z_{-b}, B, \alpha, \beta)$, 计算公式如式(3)所示:

$$p(z|z_{-b}, B, \alpha, \beta) \propto (n_z + \alpha) \frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{\left(\sum_w n_{w|z} + M\beta\right)^2} \quad (3)$$

其中: z_{-b} 表示除了词对 b 之外的所有 biterm 的主题分配, n_z 表示 biterm 分配给主题 z 的次数, $n_{w|z}$ 表示词 w 分配给主题 z 的次数, M 表示语料库中不同词的次数。一个词对 b 被分配给主题 z , 词对 b 中的两个词 w_i, w_j 也会被分配到主题 z 上。

根据词对的主题分配的次数和词共现, 全局主题分布 θ_z 和主题-词分布 $\varphi_{w|z}$ 可以进行估计:

$$\varphi_{w|z} = p(w|z) = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta} \quad (4)$$

$$\theta_z = p(z) = \frac{n_z + \alpha}{|B| + T\alpha} \quad (5)$$

其中: $\varphi_{w|z}$ 表示主题 z 中词 w 的概率, θ_z 表示主题 z 的概率, $|B|$ 表示词对总数, T 表示主题个数。

3 基于BTM的物联网服务发现

针对物联网服务的特性, 本文提出了一种基于BTM的物联网服务发现方法: 首先通过预处理物联网服务文档, 得到有效的服务特征数据集; 接着利用BTM提取服务的隐含主题, 并通过全局主题分布 θ_z 和主题-词分布 $\varphi_{w|z}$ 计算推理得到服务文档-主题概率分布 $p(z|d)$; 然后利用基于最大距离的K-means算法对服务进行聚类; 最后, 通过计算服务请求与候选服务集中服务的相似度, 返回最佳匹配结果。基于BTM的物联网服务发现框架如图3所示。

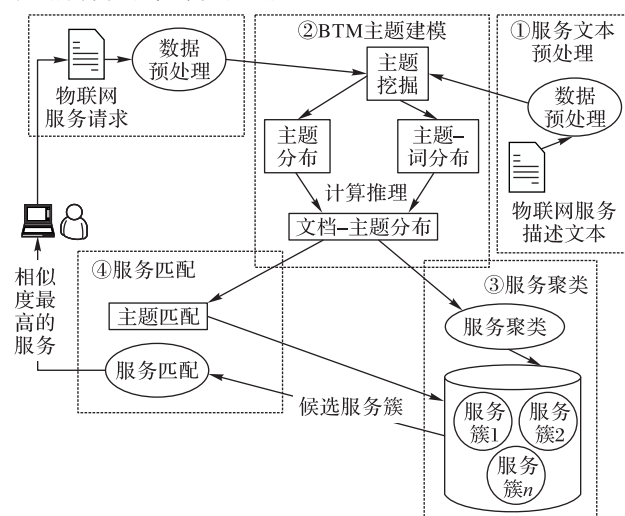


图3 基于BTM的物联网服务发现框架

Fig. 3 Service discovery framework for IoT based on BTM

3.1 数据预处理

通过特征提取、分词、去停用词、词干还原等方法, 对物联网服务描述文档进行预处理。

1) 对物联网服务描述文本中的“service name”“functional description”“interface type”等关键特征进行提取。

2) 对提取的文本中的复合词进行拆分, 如“Carbon-Dioxide”等。

3) 利用正则表达式去除一些无关词汇和符号, 如“and”“of”“&”等, 避免对建模造成影响。

4) 具有相同词干的单词具有相同的含义, 如“recommended”和“recommending”具有相同的词干“recommend”, 为了便于词语匹配, 需利用Python库NLTK的Porter Stemmer进行词干还原。

3.2 BTM建模

在BTM中, 物联网服务描述文本文档可以被看作是包含隐含主题的文档。将预处理后的数据集作为BTM的输入进行建模, 学习其隐含主题, 并通过Gibbs抽样估计出全局主题分布 θ_z 和主题-词分布 $\varphi_{w|z}$ 。由于BTM不对文档生成过程进行建模, 而是直接对语料库的词对进行建模, 我们无法直接得到服务文档-主题概率分布, 因此需要推理计算服务文档-主题分布 $p(z|d)$, 将每个物联网服务描述文本表示为隐含主题分布向量。该分布相当于文档中词对的分布和词对-主题分布的乘积, 计算公式为:

$$p(z|d) = \sum_b p(z|b) p(b|d) \quad (6)$$

其中, 以词对 b 作为中间量, 可以计算出文档 d 中词对的条件概率分布 $p(b|d)$, 计算公式如式(7)所示, $n_d(b)$ 表示词对 b 在文档 d 中出现的次数。

$$p(b|d) = n_d(b) / \sum_b n_d(b) \quad (7)$$

根据参数估计的结果, 可计算得到词对-主题分布 $p(z|b)$ 为:

$$p(z|b) = \frac{p(z) p(w_i|z) p(w_j|z)}{\sum_z p(z) p(w_i|z) p(w_j|z)} \quad (8)$$

因此, 在上述基础上, 通过全局主题分布 θ_z 和主题-词分布 $\varphi_{w|z}$ 计算推理得到服务文档-主题概率分布 $p(z|d)$, 下面给出具体的主题挖掘过程的算法。

输入 主题数 T , 数据集 $services$, 参数 α , 参数 β , 迭代次数 N ;

输出 参数 $p(z|d)$ 。

- 1) for $t=1$ to T do
- 2) Build biterm B from $services$
- 3) get each $b = (w_i, w_j)$
- 4) Initialize topic assignments randomly for all the biterms
- 5) for $i=1$ to N do
- 6) for $b \in B$ do
- 7) draw z_b from $p(z_b|B, \alpha, \beta)$
- 8) update $n_z, n_{w_i|z}, n_{w_j|z}$
- 9) compute the parameters θ_z and $\varphi_{w|z}$
- 10) compute the parameters $p(z|d)$
- 11) return $p(z|d)$

3.3 服务聚类

服务聚类是根据服务某一特征快速将服务划分成不同的集合。根据上述分析, 得到服务文档使用主题表征的向量后, 本文根据主题之间的相似度, 采用K-means算法对服务进行

聚类。

为了避免 K -means 算法在聚类时出现局部最优解问题, 本文利用余弦相似度计算 T 个主题向量两两之间的距离, 并将距离最远的两个作为初始簇中心, 在剩余的 $(T-2)$ 个主题中, 选取前面两个初始簇中心各自距离乘积最大值的那个样本点作为第三个初始簇中心, 依此类推, 可以找到 K 个初始簇中心, 进而在初始聚类中心的基础上进一步聚类。

具体的聚类算法过程如下:

步骤 1 计算 T 个服务主题两两间的相似度 $dis(n, 1)$, 选取 $dis(d_1, d_2) \geq dis(d_i, d_j) (i, j = 1, 2, \dots, T)$ 作为两个初始服务簇中心。

步骤 2 在剩余的 $T-2$ 个主题中, 选取 $dis(d_1, d_3) \times dis(d_2, d_3) \geq dis(d_1, d_i) \times dis(d_2, d_i) (i = 1, 2, \dots, K)$ 作为第三个初始簇中心, 以此类推得到 K 个服务簇中心。

步骤 3 对于剩余的 $T-K$ 个主题, 计算每个主题与 K 个服务簇中心的距离, 并将该主题聚集到距离最近的服务簇中。

步骤 4 重新计算 K 个服务簇中心。

步骤 5 重复步骤 3、4, 直到服务簇中心不再改变, 或者达到其他终止条件。

图 4 是服务聚类的具体流程。

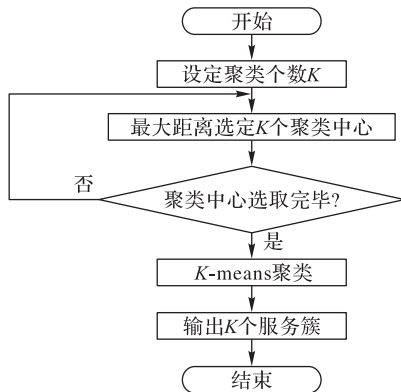


图 4 服务聚类流程

Fig. 4 Flowchart of service clustering

3.4 服务匹配

服务匹配是指根据用户的服务请求, 快速准确地匹配到与服务请求相似度最高的服务并返回。具体匹配过程如下:

1) 对于用户的服务请求, 通过 BTM 获取服务请求的隐含主题, 计算其隐含主题与各服务簇中心的距离, 将相似度最高的服务簇作为候选服务簇返回;

2) 对于候选服务簇中的每一个服务, 利用余弦相似度计算其隐含主题与服务请求的隐含主题的相似度, 并将相似度最高的服务作为匹配结果返回给用户。

4 实验分析

4.1 数据集及实验环境

本文使用采集的物联网设备服务数据集进行实验。其中, 物联网工业现场设备服务数据集包含 1 038 个服务, 涉及不同型号的射频识别 (Radio Frequency Identification, RFID) 标签、温度传感器、压力传感器、红外气体传感器、辐射传感器、电流变送器等多种物联网设备。表 1 列出了传感器的部分参数信息。

所涉及算法通过 Java、Python 编程语言实现, 实验环境为 Intel Core i5-3210M CPU @ 2.50 GHz, 内存 4.00 GB, 32 位 Windows 7 操作系统, Eclipse 及 PyCharm 开发平台。

表 1 传感器的部分参数信息

Tab. 1 Some parameters information of sensors

设备类型	部分参数
红外 二氧化碳 传感器	功能: 监测二氧化碳气体浓度并就地显示 工作温度: $0 \sim 40\text{ }^{\circ}\text{C}$ 工作湿度: $0 \sim 98\% \text{ RH}$ 供电电压: $9 \sim 25\text{ V DC}$ 输出方式: 电流输出型 $4 \sim 20\text{ mA}$
JCJ800D 压力传感器	功能: 实现对液/气体压力的测量与控制 工作温度: $-20 \sim 75\text{ }^{\circ}\text{C}$ 工作湿度: $0 \sim 100\% \text{ RH}$ 供电电压: 24 V DC 量程范围: $0 \sim 1 \sim 30\text{ MPa}$ 输出方式: 电压输出型 $0 \sim 10\text{ V}$
分体式 RS485 输出温湿度 变送器	功能: 检测温度和湿度 工作温度: $-20 \sim 55\text{ }^{\circ}\text{C}$ 工作湿度: $5 \sim 95\% \text{ RH}$ 测湿范围: $0 \sim 100\% \text{ RH}$ 测温范围: $-20 \sim 50\text{ }^{\circ}\text{C}$ 输出方式: 电流输出型 $4 \sim 20\text{ mA}$
SBWR 温度传感器	功能: 检测温度 工作温度: $-25 \sim 80\text{ }^{\circ}\text{C}$ 工作湿度: $5 \sim 95\% \text{ RH}$ 测温范围: $0 \sim 800\text{ }^{\circ}\text{C}$ 输出方式: 电流输出型 $4 \sim 20\text{ mA}$

4.2 评价指标

本文采用类内类间距离比值作为服务聚类质量的评价指标, 对于服务聚类而言, 同一类内部距离越小, 不同类之间距离越大, 表示该聚类效果明显, 计算如下:

$$R(K) = \text{within}(K) / \text{between}(K) \quad (9)$$

其中: $R(K)$ 表示类内类间距离比值, $R(K)$ 越小, 聚类质量越好。

$\text{within}(K)$ 表示类内距离, 即同一类中各服务之间的平均距离, 本文使用所有类内距离的最大值作为整个数据集的类内距离, 计算如下:

$$\text{within}(K) = \max_{1 \leq i \leq K} \left\{ \min_{1 \leq j \leq |C_i|} \left\{ \frac{1}{|C_i| - 1} \sum_{p=1, p \neq j}^{|C_i|} \|x_j - x_p\| \right\} \right\} \quad (10)$$

其中: $|C_i|$ 表示属于 C_i 类的服务的个数, x_j 和 x_p 是属于 C_i 类的服务。

$\text{between}(K)$ 表示类间距离, 即不同类最近的两个服务之间的距离, 本文使用任意两个类之间距离的最小值作为整个数据集的类间距离, 计算如下:

$$\text{between}(K) = \min_{x_p \in C_i, x_q \in C_j, i \neq j} \|x_p - x_q\| \quad (11)$$

其中: $i, j = 1, 2, \dots, K$; x_p 和 x_q 分别是属于 C_i 类和 C_j 的服务。

本文采用准确率 Precision 和归一化折损累积增益 (Normalized Discounted Cumulative Gain, NDCG) 作为服务发现结果的综合评价指标。

Precision 是指检索得到的服务中正确服务的占比, 如式

(12)所示,其中, C_i 表示第 i 个聚类簇, A 表示检索得到的服务数中正确匹配的服务数, B 表示检索得到的服务数。

$$Precision_{C_i} = |A|/|B| \quad (12)$$

NDCG用来衡量和评价检索结果算法,与服务请求相似度最高的检索结果排位越靠前,NDCG的值越大,计算如下:

$$NDCG_p = \frac{\sum_{i=1}^p \frac{2^{rel_i} - 1}{\lg(1+i)}}{\sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\lg(1+i)}} \quad (13)$$

其中: rel_i 表示在检索结果 p 个文档中第 i 个文档的相关等级,

$|REL|$ 表示这 p 个文档按照相关性从大到小的顺序排序。

4.3 实验结果与分析

对物联网服务描述文本进行特征提取、分词、词干还原等操作,生成预处理文本作为BTM建模的输入,BTM中的主题数 T 是一个经验值,根据数据集人为设定。参数 α 一般取值 $50/T$,参数 β 一般取值0.01。经过主题挖掘,可以得到服务文档-主题分布矩阵,部分数据如表2所示(结果取小数点后8位,加粗数据表示该文档中主题相关性最高的主题的概率)。

表2 服务文档-主题分布矩阵

Tab. 2 Service document-topic distribution matrix

文档	主题分布概率					
	T_1	T_2	T_3	T_4	T_5	T_6
D_1	0.000 011 54	0.000 000 04	0.018 148 69	0.000 215 00	0.872 685 27	0.000 099 07
D_2	0.000 002 00	0.000 000 73	0.000 001 41	0.999 960 86	0.000 008 64	0.000 001 01
D_3	0.000 000 06	0.161 681 17	0.774 465 46	0.000 001 87	0.000 000 19	0.024 777 79
D_4	0.000 209 45	0.000 177 40	0.000 000 59	0.000 190 97	0.033 195 96	0.964 057 00
D_5	0.000 000 60	0.719 902 11	0.000 455 16	0.000 000 34	0.279 626 18	0.000 000 25

4.3.1 BTM不同主题个数下服务聚类的比较

对具有不同主题数目的BTM在用于物联网服务聚类时的情况进行了可视化对比。在该实验中,分别设定主题数为10、15、20、30进行训练,得到的模型分别标记为BTM_10、BTM_15、BTM_20、BTM_30,部分实验结果如图5所示,图中“十字”符为各聚类中心。从实验结果图对比分析发现,随着BTM主题个数的细化,物联网服务的聚类效果发生明显的变化。计算不同主题数下各聚类数对应的类内类间距离比值,如表3所示。在 $K=7$ 时,不同主题数下类内类间距离比值较小; $T=15, K=7$ 时,类内类间距离比值最小,整体聚类效果较好(迭代次数 N 取值为2000)。

表3 不同主题数下各聚类数 K 对应的类内类间距离比值

Tab. 3 Distance ratio within and between classes under different number of topics with different K

T	K						
	5	6	7	8	9	10	11
10	0.73	0.70	0.67	0.71	0.76	0.83	0.87
15	0.68	0.60	0.53	0.62	0.69	0.75	0.76
20	0.79	0.73	0.71	0.80	0.83	0.82	0.81
30	0.86	0.82	0.79	0.81	0.84	0.87	0.87

4.3.2 不同方法下服务发现的比较

为了验证本文提出的基于BTM的物联网服务发现方法的有效性,采用4种现有服务发现研究中的常用方法TF-IDF^[14]、LDA^[15]、HDP^[10]和LDA-K^[16]作为基准与本文方法进行对比。

1)TF-IDF^[14]:该方法用TF-IDF算法计算服务之间的相似度,利用服务相似度使用 K -means方法进行服务聚类,是一种基于关键词的服务分类办法。

2)LDA^[15]:该方法建立了三层贝叶斯模型,即文档-主题-词,基于LDA模型的服务发现,通过LDA建模提取服务的隐含主题信息,并根据不同主题对服务进行分类。LDA对长文本进行主题挖掘有较好的效果,但处理短文本效果不明显。

3)HDP^[10]:一种非参数贝叶斯主题模型,可以根据数据集自动确定主题数量。与LDA模型相比,该方法更适合对具有实时性的数据集进行主题建模。

4)LDA-K^[16]:使用LDA将服务从词项空间转换到主题空间,通过建模学习得到服务的隐含主题,并使用 K -means算法对服务进行分类,该方法注重主题分布整体的相似度。

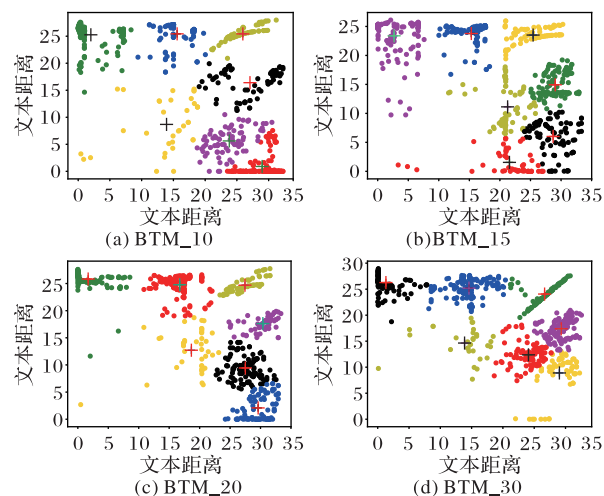


图5 不同主题个数下服务聚类效果

Fig. 5 Service clustering under different number of topics

在相同的实验环境下使用相同的数据集,对以上方法提出相同的服务请求进行查询。图6和图7分别给出了在不同数量的服务查询下5种方法的Precision和NDCG指标的评估结果。图6是5种方法的物联网服务查询准确率的对比,可以看出,随着服务查询数量的增长,所有方法查准率都呈下降趋势,查询数量为35时趋于稳定,本文方法整体上要优于其他方法。通过计算可知,基于BTM的服务查询平均准确率相比TF-IDF、LDA、HDP和LDA-K方法分别提高了5.91%、4.01%、2.92%、2.46%。

图7是5种方法在不同查询数量下的NDCG值对比,NDCG值越大说明与服务请求相似度最高的检索结果排位越靠前。从图7中可以看出,5种方法的NDCG值随着服务查询数量的增加均有所下降,相比其他4种常用方法,本文方法的查询结果相似度更高。针对短文本的物联网服务,本文方法可以直接建模学习,无需进行文本扩充,有效解决了数据稀疏性问题,服务匹配返回最佳服务的效果得到了提升。

综合分析图6和图7的结果可以看出:TF-IDF侧重通过计算文本中单词的词频和逆文档频率进行服务聚类;LDA、HDP和LDA-K主要针对于较长文本进行文档级建模;而BTM主题模型则是对词共现进行建模来增强主题挖掘,利用整个

语料库的聚合模式来学习文本的隐含主题,有效地解决了短文档级的数据稀疏问题。因此,基于 Precision 和 NDCG 的综合分析,对文本较短、语义性差的物联网服务进行聚类,BTM 有效地提高了服务发现的效果。

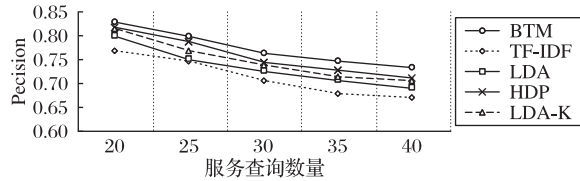


图 6 5种方法的物联网服务查询准确率对比

Fig. 6 Precision comparison of IoT service query by five methods

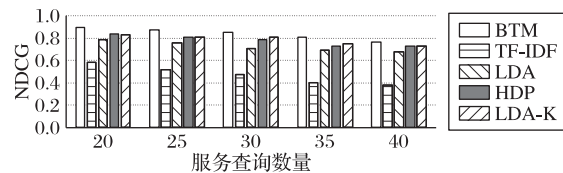


图 7 5种方法在不同查询数量下的NDCG值对比

Fig. 7 NDCG comparison of five methods under different service queries

5 结语

如何快速有效地找到最佳服务以满足不断增长的服务需求是当前物联网应用中需要解决的关键问题之一,由于物联网服务文本短、语义性差等特性,应用现有的服务发现方法不能很好地匹配到最佳服务。针对这一问题,本文提出了一种基于 BTM 的物联网服务发现方法,该方法利用 BTM 挖掘物联网服务的隐含主题,对服务进行聚类并返回服务请求的最佳匹配结果。实验结果分析表明,该方法在 Precision 和 NDCG 方面比常用的其他方法均有更好的效果。

本文的研究主要集中于物联网服务属于短文本、语义性差这一特性上,今后,我们将会在本文的基础上进一步考虑为物联网服务添加语义标记、服务的实时变化性等特性对物联网服务发现的影响。

参考文献 (References)

- [1] 中国信息通信研究院. 物联网白皮书(2018年)[EB/OL]. [2019-04-25]. http://www.caict.ac.cn/kxyj/qwfb/bps/201812/t20181210_190297.htm. (China Academy of Information and Communications Technology. White Paper on Internet of Things (2018) [EB/OL]. [2019-04-25]. http://www.caict.ac.cn/kxyj/qwfb/bps/201812/t20181210_190297.htm.)
- [2] GUINARD D, TRIFA V, KARNOUSKOS S, et al. Interacting with the SOA-based Internet of things: discovery, query, selection, and on-demand provisioning of Web services[J]. IEEE Transactions on Services Computing, 2010, 3(3): 223-235.
- [3] LIU Y, DU F, SUN J, et al. A crowdsourcing-based topic model for service matchmaking in Internet of things[J]. Future Generation Computer Systems, 2018, 87: 186-197.
- [4] JIANG B, YE L, WANG J, et al. A semantic-based approach to service clustering from service documents[C]// Proceedings of the 2017 IEEE International Conference on Services Computing. Piscataway: IEEE, 2017: 265-272.
- [5] 李平. 面向物联网设备的自动语义标注和服务封装的研究与应用[D]. 广州: 华南理工大学, 2017: 6-8. (LI P. Research and application of IoT device-oriented automatic semantic annotations and service encapsulation[D]. Guangzhou: South China University

of Technology, 2017: 6-8.)

- [6] 王书龙, 侯义斌, 高放, 等. 基于本体的物联网设备资源描述模型[J]. 北京工业大学学报, 2017, 43(5): 762-769. (WANG S L, HOU Y B, GAO F, et al. Ontology-based resource description model for things in IoT[J]. Journal of Beijing University of Technology, 2017, 43(5): 762-769.)
- [7] SANTOS A C, PEDROSA L D, KUIPERS M, et al. Resource description language: a unified description language for network embedded resources[J]. International Journal of Distributed Sensor Networks, 2012, 2012: 860-864.
- [8] TEIXEIRA T, HACHEM S, ISSARNY V. Service oriented middleware for the internet of things: a perspective[C]// Proceedings of the 2011 European Conference on a Service based Internet, LNCS 6994. Berlin: Springer, 2011: 220-229.
- [9] CASSAR G, BARNAGHI P, WANG W, et al. A hybrid semantic matchmaker for IoT services[C]// Proceedings of the 2012 IEEE International Conference on Green Computing and Communications. Piscataway: IEEE, 2012: 210-216.
- [10] 魏强, 金芝, 许焱. 基于概率主题模型的物联网服务发现[J]. 软件学报, 2014, 25(8): 1640-1658. (WEI Q, JIN Z, XU Y. Service discovery for Internet of things based on probabilistic topic model[J]. Journal of Software, 2014, 25(8): 1640-1658.)
- [11] 肖巧翔, 曹步清, 张祥平, 等. 基于 Word2Vec 和 LDA 主题模型的 Web 服务聚类方法[J]. 中南大学学报(自然科学版), 2018, 49(12): 2979-2985. (XIAO Q X, CAO B Q, ZHANG X P, et al. Web services clustering based on Word2Vec and LDA topic model[J]. Journal of Central South University (Science and Technology), 2018, 49(12): 2979-2985.)
- [12] YAN X, GUO J, LAN Y, et al. A bitern topic model for short texts[C]// Proceedings of the 22nd International Conference on World Wide Web. New York: ACM, 2013: 1445-1456.
- [13] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. PNAS, 2004, 101(S1): 5228-5235.
- [14] ZHANG Y, GONG L, WANG Y. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang University-SCIENCE A, 2005, 6(1): 49-55.
- [15] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [16] CAO B, LIU X, LI B, et al. Mashup service clustering based on an integration of service content and network via exploiting a two-level topic model[C]// Proceedings of the 2016 IEEE International Conference on Web Services. Piscataway: IEEE, 2016: 212-219.

This work is partially supported by the National Key R&D Progamme of China (Research and Development Program — "cyberspace security" (2018YFB0803402)).

WANG Shuman, born in 1994, M. S. candidate. Her research interests include service computing, information processing.

LI Aiping, born in 1974, Ph. D., professor. Her research interests include IoT data processing, wireless sensor network.

DUAN Ligu, born in 1970, Ph. D., associate professor. His research interests include Chinese information processing, knowledge mapping.

FU Jia, born in 1995, M. S. candidate. Her research interests include service computing, information processing.

CHEN Yongle, born in 1983, Ph. D., associate professor. His research interests include wireless sensor network, IoT security.