相似度自适应估计的物联网实体高效搜索方法

张普宁 亢旭源* 刘宇哲 李学芳 吴大鹏 王汝言

(重庆邮电大学通信与信息工程学院 重庆 400065)

(重庆高校市级光通信与网络重点实验室 重庆 400065)

(泛在感知与互联重庆市重点实验室 重庆 400065)

摘 要：针对现有相似实体搜索方法缺乏对于观测序列长度的自适应性，且搜索过程数据存储开销过大，搜索结果准确性较低的问题，该文提出相似度自适应估计的物联网实体高效搜索方法(SAEES)。首先，设计了轻量级观测序列分段表示方法，对传感器采集的实体原始观测序列进行轻量级分段压缩表示，以降低实体观测序列的存储开销。然后，提出了观测序列相似度自适应估计方法，实现对不同观测序列长度的实体相似性的准确估计。最后，设计了高效的相似实体搜索匹配方法，依据所估计的实体相似度进行实体的准确搜索匹配。仿真结果表明，所提方法可大幅提高相似实体搜索的效率。

关键词：物联网；实体搜索；相似度；自适应估计

中图分类号：TN915; TP393 **文献标识码**：A **文章编号**：1009-5896(2020)07-1702-08

DOI: 10.11999/JEIT190541

Efficient Search Method for IoT Entities with Similarity Adaptive Estimation

ZHANG Puning KANG Xuyuan LIU Yuzhe LI Xuefang
WU Dapeng WANG Ruyan

(*School of Communication and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

(*Optical Communication and Networks Key Laboratory of Chongqing*, *Chongqing* 400065, *China*)

(*Ubiquitous Sensing and Networking Key Laboratory of Chongqing*, *Chongqing* 400065, *China*)

Abstract: The existing similar entity search method has poor adaptability to the length of the observed sequence, and the data storage overhead in the search process is too large, and the accuracy of the search result is insufficient. To this end, an efficient search method is proposed for the IoT Entity Search with Similarity Adaptive Estimation (SAEES). Firstly, in order to reduce the storage overhead of the entity observation sequence, a lightweight method of segmentation representation of the observation sequence is designed to perform a lightweight segmentation compression representation of the original observation sequence of the entity collected by the sensor. Then, in order to achieve an accurate estimation of the similarity of entities with different observation sequence lengths, an adaptive estimation method for observation sequence similarity is proposed. Finally, by exploiting the designed efficient similar entity search matching method, the exact search matching of the entity is completed according to the estimated entity similarity. The simulation results show that the proposed method can greatly improve the efficiency of similar entity search.

Key words: Internet of Things (IoT); Entity search; Similarity; Adaptive estimation

收稿日期：2019-07-18；改回日期：2020-03-07；网络出版：2020-04-11

*通信作者： 亢旭源 kangxuyuan163@163.com

基金项目：国家自然科学基金(61871062, 61901071)，重庆市高校创新团队建设计划资助项目(CXTDX201601020)，重庆市自然科学基金面上项目(cstc2019jcyj-msxmX0303)，重庆市教委科学技术研究项目(KJQN201800615)，第五批重庆市高校优秀人才支持计划(渝教人发[2017]29号)

Foundation Items: The National Natural Science Foundation (61871062, 61901071), The Program for Innovation Team Building at Institutions of Higher Education in Chongqing (CXTDX201601020), The General Project of Natural Science Foundation of Chongqing (cstc2019jcyj-msxmX0303), The Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN201800615), The Fifth Supporting Plan for Chongqing's University Excellent Talents (Chongqing Municipal Education Commission, No.29 [2017])

1　引言

随着物理空间[1]与网络空间融合的逐步深化，用户对获取物理实体信息的实时性、有效性、可靠性[2]需求日益提高，推动了物联网搜索技术的诞生[3]。物联网搜索技术从物联网中获取实体信息(如物体、人等)，并对获取到的信息进行有组织、有序地管理与存储[4,5]，以方便用户进行搜索[6]。实体状态信息由传感器感知并生成，用户可通过指定需求状态对实体进行搜索。例如，查询当前办公楼里温度适宜、空闲的会议室，搜索当前安静、人数较少的餐厅等。搜索的内容既包含会议室、餐厅等实体的静态属性，也涵盖当前、附近等动态时空属性，及温度适宜、空闲等时变状态信息[7,8]。传统的互联网搜索方法仅面向静态的虚拟信息资源，不适用于对状态动态变化的物理实体的搜索[9]。

目前，已有的物联网实体搜索原型系统的设计借鉴了传统的互联网搜索模式[10]。MAX[11]通过将物理实体的描述信息存储于实体关联传感设备，用户通过指定实体描述关键字来搜寻物理实体。Snoogle[12]通过将搜索用户输入的关键字，与附着在物理实体上的传感器中存储的实体静态关键字描述进行匹配，以搜索与其相关的前k个实体。Microsearch[13]首先对存储在传感器中的实体文本描述信息进行索引，然后，采用TF-IDF算法对实体描述信息的匹配度进行评分与排名，将与查询约束相关的前k个实体返回给用户。

上述搜索系统均基于实体关键字描述[14]来搜寻匹配实体。然而，用户间的主观认知具有较大的差异性，机器显然缺乏对与用户主观认知相关信息的解析能力(机器对于人少、安静等实体状态缺乏统一的认知度量标准)。“以图搜图”的互联网搜索模式[15]，启发了物联网“以数据搜数据”的新型搜索模式。例如，用户在搜索气候适宜的旅游度假地时，可通过预先记录符合用户舒适度需求的相关传感设备读数，搜索与之具相似观测序列输出的实体，从而，避免了用户主观认知对实体搜索结果的影响。

当前，关于前述相似实体搜索的研究较为匮乏，已有的研究中，文献[16]通过对实体的观测序列构造模糊集，计算候选实体的观测序列与用户指定观测序列间的相似度，从而，获得具有相似输出的实体集合。文献[17]提出物联网中基于相似性的实体搜索方法-PLSS，通过计算指定实体与候选实体间观测序列的欧式距离来估计其相似度，从而，依据相似度进行匹配实体的搜寻与排序。然而，上述研究需候选实体与指定实体的观测序列长度一致，且在时间维度上严格对齐，无法适用于因采样频率或噪声等原因，而造成观测序列长度不一的实体间的相似度估计。并且，物联网中实体的数量众多，存储全部实体所有时刻的观测序列将带来严重的存储资源开销[18]。

为解决上述问题，本文提出了相似度自适应估计的物联网实体高效搜索(Similarity Adaptive Estimation IoT Entity Search, SAEES)方法。主要贡献如下：(1)设计轻量级观测序列分段表示方法，在降低观测序列存储开销的基础上，减小网关的计算开销；(2)提出观测序列相似度自适应估计方法，以自适应评估具不同观测序列长度的实体间的相似度；(3)设计相似实体高效搜索方法，基于所估计的实体相似度，高效匹配满足用户搜索需求的实体集合。

2　轻量级观测序列分段表示方法

物联网中智慧实体的数量众多且其状态时变性较强[19,20]，存储实体观测序列原始数值，将极大地消耗存储资源，因此，需对待搜实体的观测序列进行压缩表示，以降低存储开销。本文提出一种轻量级在线分段数据表示方法，在满足物联网搜索实时性要求基础上，实现对海量实体观测序列的在线压缩表示。

2.1　在线分段方法

假设传感器在t_i时刻采集的数据为s_i，则实体观测序列可表示为$(t_1,s_1),(t_2,s_2),\cdots,(t_i,s_i),\cdots,(t_n,s_n)$。寻求与包含$n$个数据的数据集$(T,S)$拟合程度最高的线性回归方程定义为

$$\hat{s}=kt+b,\quad t\in t_1,t_2,\cdots,t_n \tag{1}$$

其中，$\hat{s}_i$表示传感器在t_i时刻的观测值s_i的近似值，b为待定常数，k为回归系数。

寻求最佳线性回归方程的过程，等价于求解使得残差平方和达到最小值的(k,b)组合的过程。定义残差平方和如式(2)

$$Q=\sum_{i=1}^{n}(s_i-\hat{s}_i)^2 \tag{2}$$

为求解线性回归方程(1)，需确定k与b的值。本文采用最小二乘法求解k, b的值，计算方法如下

$$k=\frac{\sum_{i=1}^{n}(t_i-\bar{T})(s_i-\bar{S})}{\sum_{i=1}^{n}(t_i-\bar{T})^2} \tag{3}$$

$$b=\bar{S}-k\bar{T} \tag{4}$$

其中，$\bar{T}$为t_i的均值，$\bar{S}$为s_i的均值。将式(1)，式(3)，式(4)代入式(2)化简可得

$$Q=\sum_{i=1}^{n}(s_i-\bar{S})^2-\frac{\left[\sum_{i=1}^{n}(t_i-\bar{T})(s_i-\bar{S})\right]^2}{\sum_{i=1}^{n}(t_i-\bar{T})^2} \tag{5}$$

为便于式(5)计算，令$L_{tt}=\sum_{i=1}^{n}(t_i-\bar{T})^2$，$L_{ts}=\sum_{i=1}^{n}(t_i-\bar{T})(s_i-\bar{S})$，$L_{ss}=\sum_{i=1}^{n}(s_i-\bar{S})^2$，其中，$\bar{T}=\left(\sum_{i=1}^{n}t_i\right)/n$，$\bar{S}=\left(\sum_{i=1}^{n}s_i\right)/n$，则式(3)与式(5)可进一步表示为

$$k=L_{ts}/L_{tt} \tag{6}$$

$$Q=L_{ss}-L^2{}_{ts}/L_{tt} \tag{7}$$

由上可知，通过计算$\bar{T}$, $\bar{S}$, L_{tt}, L_{ts}, L_{ss}，即可求得观测序列$(t_1,s_1),(t_2,s_2),\cdots,(t_i,s_i),\cdots,(t_n,s_n)$的线性回归方程。

当搜索系统接收到新的实体观测数据后，为求解此时新观测序列的线性回归方程，需重新计算$\bar{T}$, $\bar{S}$, L_{tt}, L_{ts}, L_{ss}。为降低因重复计算而带来的计算开销，本文设计了面向流数据的线性回归方程递推方法，其原理为：基于原有统计量$\bar{T}$, $\bar{S}$, L_{tt}, L_{ts}, L_{ss}，以及新到达数据(t_{n+1},s_{n+1})，计算如上统计量的更新量。由此，即可提高计算面向流数据的线性回归方程的效率。$\bar{T}$, $\bar{S}$, L_{tt}, L_{ts}和L_{ss}的递推方法为

$$\left.\begin{aligned}
\bar{S}_{n+1}&=\frac{n\times\bar{S}_a}{n+1}+\frac{s_{n+1}}{n+1}\\
\bar{T}_{n+1}&=\frac{n\times\bar{T}}{n+1}+\frac{t_{n+1}}{n+1}\\
L_{(tt)_{n+1}}&=L_{tt}+n\times\frac{[\bar{T}-(t_{n+1})]^2}{n+1}\\
L_{(ts)_{n+1}}&=L_{ts}+n\times\frac{[\bar{T}-(t_{n+1})][\bar{S}-(s_{n+1})]}{n+1}\\
L_{(ss)_{n+1}}&=L_{ss}+n\times\frac{[(\bar{S}-s_{n+1})]^2}{n+1}
\end{aligned}\right\} \tag{8}$$

其中，$\bar{T}_{n+1}$, $\bar{S}_{n+1}$, $L_{(tt)_{n+1}}$, $L_{(ts)_{n+1}}$和$L_{(ss)_{n+1}}$分别表示新数据(t_{n+1},s_{n+1})到达后，$\bar{T}$, $\bar{S}$, L_{tt}, L_{ts}, L_{ss}的更新值。在求解各统计量的递推式后，传感器每采集一个新的数据点，则以递推式计算出该分段在增加该数据点之后的线性回归方程。

2.2 分段点决策方法

观测序列分段点的确定对保障线性回归方程表征实体观测序列$(t_1,s_1),(t_2,s_2),\cdots,(t_i,s_i),\cdots,(t_n,s_n)$的精度至关重要。在实体观测序列分段过程中，阈值$\delta$约束了对观测序列进行线性表示的误差容忍范围。不同阈值δ的选取直接决定了分段线性表示方法的精度。本文拟采用残差平方和$Q=\sum_{i=1}^{n}(s_i-\hat{s}_i)^2$作为判断以某个数据点为分段点，对相邻分段点间的观测数据点进行线性表示的合理性的度量依据。同理，每次接收新数据后对应更新Q将带来极大的计算开销，因此，为减少重复计算开销，利用前述所得递推式(8)结合式(7)可得到残差平方和递推式定义为

$$Q=L_{(ss)_{n+1}}-L^2{}_{(ts)_{n+1}}/L_{(tt)_{n+1}} \tag{9}$$

本文采用$\sqrt{(Q/n)}/\bar{T}$作为判断分段的标准，其表示残差平方和Q在时间序列上的量化比例，直观地显示了新数据点加入后，原始观测值与分段后拟合值之间的偏差。当有新数据点到达之后，递推更新$\bar{T}$, $\bar{S}$, L_{tt}, L_{ts}与L_{ss}的取值，进而，判断偏差$\sqrt{(Q/n)}/\bar{T}$是否超过阈值δ。若超过，则将该数据点之前的数据点集合作为一个分段；若未超过阈值δ，则将该点添加到当前分段中。

3 观测序列相似度自适应估计方法

考虑到传感器由于采集频率或噪声等原因可能会造成观测数据的缺失，并且在经过轻量级数据表示后，任意两实体的分段序列长度极有可能不同。因此，本文设计了可自适应观测序列长度的实体相似度估计方法，计算不同实体间任意长度分段观测序列的相似度。

如前所述，传感器的原始观测序列经轻量级数据表示方法处理之后，以多个线性回归方程的有序集合形式进行表达。假定分段序列集合为$V=\{v_1,v_2,\cdots,v_i,\cdots,v_p\}$，则第$i$个分段的线性回归方程为$v_i=k_i't+b_i'$。每个分段的线性回归方程可由回归系数$(k',b')$唯一确定，则用户提交的实体观测序列可采用前述轻量级数据表示方法转化为分段观测序列$S=\{s_1,s_2,\cdots,s_j,\cdots,s_q\}$，$s_1=(k_1,b_1)$, $s_2=(k_2,b_2),\cdots,s_j=(k_j,b_j),\cdots,s_q=(k_q,b_q)$。

假设在同一观测时段τ内，有分属两实体的分段序列长度为p的$V=\{v_1,v_2,\cdots,v_i,\cdots,v_p\}$，长度为$q$的$S=\{s_1,s_2,\cdots,s_j,\cdots,s_q\}$。为估计两序列间的相似度，首先计算分段序列$V$与$S$中各元素之间的距离，生成距离矩阵$\boldsymbol{D}_{p\times q}$，如式(10)所示

$$\boldsymbol{D}_{p\times q}=\begin{bmatrix}
D(v_1,s_1) & D(v_1,s_2) & \cdots & D(v_1,s_j) & \cdots & D(v_1,s_q)\\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\
D(v_i,s_1) & D(v_i,s_2) & \cdots & D(v_i,s_j) & \cdots & D(v_i,s_q)\\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\
D(v_p,s_1) & D(v_p,s_2) & \cdots & D(v_p,s_j) & \cdots & D(v_p,s_q)
\end{bmatrix} \tag{10}$$

其中，$D(v_i,s_j)=\|v_i-s_j\|$，表示v_i与s_j之间的欧式距离，计算公式如式(11)

$$D(v_i,s_j)=\sqrt{(k'_i-k_j)+(b'_i-b_j)} \tag{11}$$

若$p=q$，表明两分段序列长度相等，则根据式(11)可直接计算两分段序列之间的相似度。若$p\neq q$，则应将两分段序列在长度上规整对齐。本文采用动态规整方法实现非等长序列的对齐，进而，计算其相似度$D(V,S)$。

首先，需在两观测序列间的距离矩阵$\boldsymbol{D}_{p\times q}$的有效路径中搜寻最优路径$W_{\text{best}}=\{w_1,w_2,\cdots,w_r,\cdots,w_K\}$，$\max(p,q)\le K<p+q-1$，以使分段序列$V$与$S$之间的累积距离最小。$W_{\text{best}}$的第$r$个元素定义为$w_r=(v_i,s_j)_r$，表示$v_i$与$s_j$之间的映射关系。有效的规整路径需满足以下条件

$$\left.\begin{array}{l} w_1=(v_1,s_1),w_k=(v_p,s_q)\\ w_r=(v_i,s_j),w_{r+1}=(v_{i'},s_{j'}),\\ \forall i\le i'\le i+1,j\le j'\le j+1 \end{array}\right\} \tag{12}$$

式(12)表示：(1)规整路径W从$w_1=(v_1,s_1)$开始，至$w_r=(v_p,s_q)$结束；(2)$w_r=(v_i,s_j)$，$w_{r+1}=(v_{i'},s_{j'})$满足$i\le i'\le i+1$，$j\le j'\le j+1$，以保证分段按序映射。如前所述，最优路径定义为使得V与S间累积距离最小的路径，如式(13)所示

$$\text{DTW}(V,S)=\min\left\{\frac{1}{K}\sum_{k=1}^{K}D(w_K)\right\} \tag{13}$$

其中，分母中的K主要是用来对不同的长度的规整路径做补偿。因为不同的路径其长短不同，较长的路径在两观测序列间存在有较多的“点对”，会有较多的距离累加上去，所以总距离除以K得到单位路径的距离。

为求解式(13)，采用动态规整方法构造累积距离矩阵$\boldsymbol{\gamma}$，如式(14)所示

$$\boldsymbol{\gamma}_{p\times q}=\begin{bmatrix} \gamma(v_1,s_1) & \gamma(v_1,s_2) & \cdots & \gamma(v_1,s_j) & \cdots & \gamma(v_1,s_q)\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ \gamma(v_i,s_1) & \gamma(v_i,s_2) & \cdots & \gamma(v_i,s_j) & \cdots & \gamma(v_i,s_q)\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ \gamma(v_p,s_1) & \gamma(v_p,s_2) & \cdots & \gamma(v_p,s_j) & \cdots & \gamma(v_p,s_q) \end{bmatrix} \tag{14}$$

$$\gamma(v_i,s_j)=D(v_i,s_j)+\min\left\{\begin{array}{l} D(v_{i-1},s_j)\\ D(v_{i-1},s_{j-1})\\ D(v_i,s_{j-1}) \end{array}\right. \tag{15}$$

其中，$\gamma(v_i,s_j)$表示累积距离，为v_i与s_j之间的距离$D(v_i,s_j)$与可到达$D(v_i,s_j)$的最小的邻近元素的累积距离之和，则$\gamma(v_p,s_q)$为V与S间的最小累积距离。因此据式(13)，两观测序列间的DTW距离可表示为

$$\text{DTW}(V,S)=\gamma(v_p,s_q),(p\neq q) \tag{16}$$

结合式(11)与式(16)，将实体观测序列间的相似距离转化为相似性度量，则其相似度如式(17)所示

$$\text{Sim}(V,S)=\left\{\begin{array}{ll} \dfrac{1}{D(V,S)}, & p=q\\ \dfrac{1}{\gamma(v_p,s_q)}, & p\neq q \end{array}\right. \tag{17}$$

其中，$D(V,S)=D(v_1,s_1)+D(v_2,s_2)+\cdots+D(v_p,s_q)$，$\text{Sim}(V,S)$表示实体观测序列间的相似度。由此，搜索系统即可自适应估计不同实体的两任意长度观测序列间的相似度。

4　相似物联网实体高效搜索方法

遍历搜索方法将给搜索系统带来严重的通信开销，为此，基于前述轻量级数据表示方法与相似度估计方法，本部分提出了高效的相似实体搜索方法，依据所估计实体间的相似度进行实体的高效搜寻匹配。

如图1所示，所提相似实体搜索架构由客户端、服务器、IoT网关、传感器与实体构成。用户通过客户端提交指定的基准实体观测序列；服务器负责响应用户的搜索请求，并根据搜索的内容将搜索请求发布至相应的IoT网关，同时服务器也负责将匹配用户请求的实体列表返回给用户；IoT网关负责对传感器上传的原始数据进行轻量级分段表示，并负责将其与用户指定的实体观测序列进行相似度估计与匹配；传感器负责采集关联实体状态数据，并将其上传到IoT网关。具体搜索过程定义如下：(1)传感器周期性观测实体状态数据并上报至IoT网关；(2)IoT网关采用所提轻量级分段表示方法，对传感器的上报数据进行分段线性表示，并存储分段表示后的数据；(3)用户由客户端向服务器发送搜索请求，提交搜索请求内容中包含指定时段的基准实体的观测序列；(4)服务器将用户的搜索请求发布到相应的IoT网关；(5)IoT网关接收到服务器发布的搜索请求后，采用所提分段表示方法与相似度估计方法计算候选实体的观测序列与基准观测序列的相似度；(6)IoT网关选取具有最高相似度的前k个实体上报给服务器；(7)服务器将与用户搜索请求匹配的搜索结果返回给用户。

5　仿真分析

本文采用Intel Berkeley数据集[21]中54个传感器节点所采集的温度数据，对本文所提的轻量级数据表示方法，以及相似实体搜索方法的性能进行仿真验证。实验参数设置如下：用户指定传感器个数为21个，搜索范围为54个传感器节点，每个传感器节点数据量为10000个数据点，传感器采样间隔为31 s。

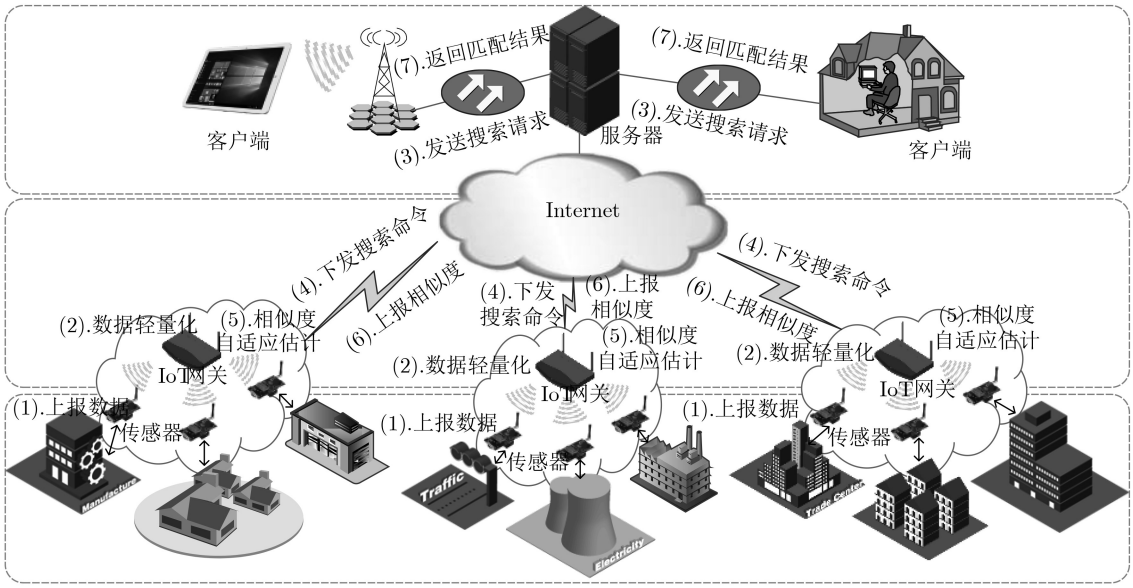

图 1 相似实体搜索架构设计

搜索方式是通过设定相似度阈值区间$[0, \delta]$，搜索满足相似度阈值区间$[0, \delta]$的传感器节点，相似度阈值可动态调整。搜索结果均为执行10000次搜索后所得的平均值。仿真环境为64位Windows7操作系统，处理器为Intel(R) Core(TM) i5-4210M, 4 GB内存，CPU 2.6 GHz，仿真软件为Matlab R2018b。

5.1 轻量级分段表示方法性能评估

该部分对所提轻量级分段表示方法进行了性能验证，并与文献[15]中PLSS算法进行了对比。其中，PLSS算法思想是通过预先设定分段阈值，在阈值内执行分段，分段结束后，通过候选实体观测序列的拟合值与指定实体观测序列的原始数据间的欧氏距离确定相似度，进而依据相似度进行搜索以及排序。

图2仿真设置在不同分段数量情况下，对比两种方法的拟合误差。从图2可以看出，所提SAEES算法与PLSS算法随着分段数量的增多，两者的分段拟合误差均呈现出下降的趋势，这是因为分段数量增多，各分段方程对原始观测序列数据的拟合度提高，则分段方程相对于原始观测序列数据的拟合误差降低。然而，SAEES算法的分段拟合误差整体要小于PLSS算法。这是因为SAEES算法的分段点确定方法较PLSS算法更准确，分段的方程也随着新数据点的加入而及时调整，进而在整体上提高了对原始数据的拟合度。而PLSS算法是预先设定误差阈值，在误差阈值范围内执行分段。于是，在给定分段数的情况下，PLSS算法的拟合误差曲线变化较为明显。

为验证SAEES算法与PLSS算法的数据处理效率，本部分验证了在输入不同数据量的情况下，两者所消耗的时间，其仿真结果如图3所示。由图3可知，随着两种算法处理数据量的不断增加，两者的时间消耗均呈现上升趋势，这是因为数据量的增加加大了计算机处理任务量，从而两者时间消耗增加，但是PLSS算法随着处理数据量的增多，时间消耗上升趋势明显大于SAEES算法。这是由于PLSS算法在分段过程中，对每一个新到达的数据点判断是否为分段点时，都需要重新计算当前段内所有数

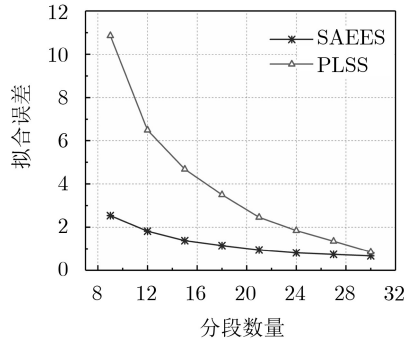

图 2 不同分段数量下的拟合误差

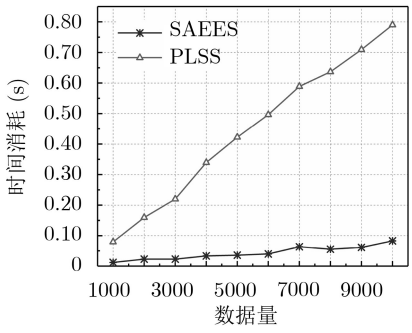

图 3 不同数据量下的分段时间消耗

据生成的线性回归表达式及拟合误差，存在大量的重复计算，而SAEES算法通过构建递推式，可在新数据点到达后直接递推计算段内数据点的线性回归表达式以及拟合误差，从而避免了重复计算所产生的时间消耗。因此，对于物联网环境中海量流数据而言，本文所提的SAEES算法在数据的轻量化表示方面更具优势。

5.2 相似实体搜索方法性能验证

该部分对本文所提方法，分别在不同相似度阈值、不同分段数量、不同候选实体观测序列长度及不同基准实体观测序列长度下进行了查准率性能的验证。

由图4可知，随着相似度阈值的增大，SAEES算法与PLSS算法的查准率均呈现出下降趋势，这是因为随着相似度阈值增大会将更多的实体置于阈值区间内，搜索结果中实体数量增加，进而导致查准率下降，但PLSS算法的查准率较SAEES算法下降速度更快，这是因为PLSS算法采用分段后观测序列的拟合值与原始数据的绝对差值作为相似距离，由此计算出的实体观测序列间相似距离数值差异较小，导致相似度的区分度较低。而本文所提方法在对齐不同长度的分段序列基础上，利用各分段的回归系数计算相似度，由此估计的相似度区分度更高，因此，最终的查准率更高。相较于PLSS算法，SAEES算法在不同相似度阈值下的查准率方面平均提升了约43.9%。

图5验证了所提搜索方法在不同分段数下的性能。由图5可知，SAEES算法随着分段数的增加，查准率呈上升趋势，而PLSS算法的查准率变化幅度较小，这是因为PLSS算法采用分段后观测序列的拟合值与原始数据的绝对差值计算相似度，所以分段数量对查准率的影响较弱。而SAEES算法基于DTW模型计算分段后观测序列的相似度时需将不同长度的分段观测序列规整对齐，同时分段观测序列的长度取决于分段数量，于是分段数量对搜索系统的查准率影响较大。由图可看出，SAEES算法在此种情况下的查准率远大于PLSS算法，性能平均提升了约33.6%。

图6验证了在不同候选实体观测序列长度条件下所提方法的查准率性能。由图6可见，随着候选实体观测序列长度的增加，两种算法的查准率都呈现出逐渐上升的趋势，这是因为观测序列长度的增加，观测序列间相似度计算的准确性提高，所以查准率均呈现上升趋势。从图中可以看出，候选实体观测序列长度的变化对PLSS算法查准率的影响更加明显，这是因为PLSS算法的相似度计算方式仅利用了分段后观测序列的拟合值与原始数据的绝对差值，而未考虑观测序列的变化趋势，所以观测序列长度的大小对搜索系统的查准率影响较大。而SAEES算法充分考虑了观测序列的变化趋势，因此影响较弱。另外，SAEES算法较PLSS算法在不同候选实体观测序列长度的情况下的查准率平均提升了约37.7%。

图7验证了在用户提交的不同基准实体观测序列长度情况下所提方法的性能。用户通过指定基准实体观测序列发起搜索，与给定序列长度的候选实体观测序列进行相似度匹配。由图7可知，SAEES算法与PLSS算法的查准率，随基准实体观测序列长度的增加呈现上升趋势，这是因为基准实体观测序列长度的增加提高了观测序列间相似度计算的准确性，故而查准率均呈现上升趋势，但用户提交的

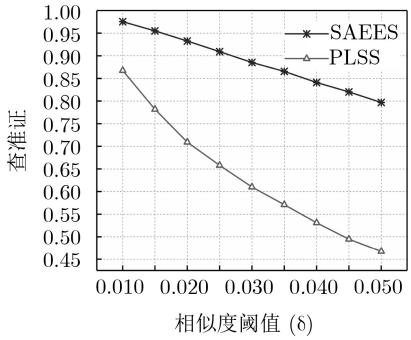

图 4 不同相似度阈值下传感器的查准率

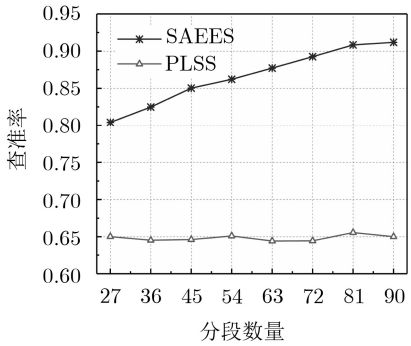

图 5 不同分段数下的查准率

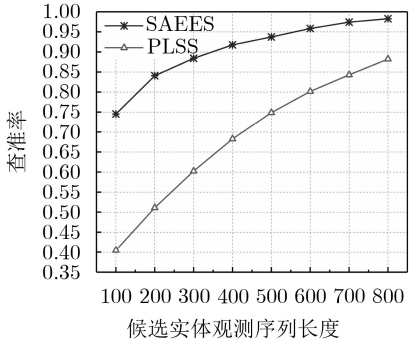

图 6 不同候选实体观测序列长度下的查准率

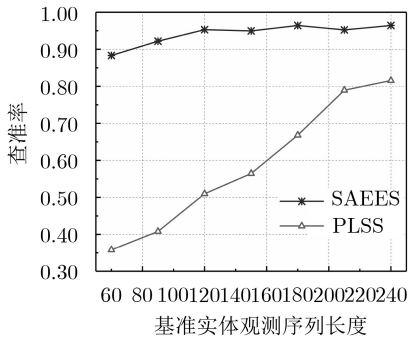

图 7 不同基准实体观测序列长度下的查准率

基准实体观测序列长度的大小对PLSS算法影响更大。这是由二者算法的相似度计算方式不同导致的，本文所提的相似度计算方法充分考虑了观测序列的变化趋势并且能自适应规整观测序列长度，因此用户提交的基准实体观测序列长度的大小对搜索系统的查准率影响较小。同时，SAEES算法相对PLSS算法在此情况下性能平均提升了72.9%。

6 结束语

本文提出了适用于物联网的相似实体高效搜索方法，设计了轻量级观测序列分段表示方法对实体原始观测序列进行近似表示，提出了实体观测序列长度自适应的相似度估计方法，通过自适应规整实体间观测序列的长度，以提高不同实体间相似度计算的精度，进而，提出高效的相似实体搜索方法，基于所估计的实体相似度，进行物联网实体的高效搜索匹配。仿真结果表明，所提方法在降低实体观测序列的存储开销与计算开销的同时，可有效提高物联网相似实体的搜索效率。有关下一步的研究工作，我们将关注传感器观测数据的多维度属性对实体搜索过程的影响，采用深度学习方法通过搜索场景动态计算不同维度属性的影响因子，从而进一步提高物联网实体的准确度。

参 考 文 献

[1] WU Dapeng, SHI Hang, WANG Honggang, *et al*. A feature-based learning system for internet of things applications[J]. *IEEE Internet of Things Journal*, 2019, 6(2): 1928–1937. doi: 10.1109/JIOT.2018.2884485.

[2] ZHANG Puning and MA Jie. Channel characteristic aware privacy protection mechanism in WBAN[J]. *Sensors*, 2018, 18(8): 2403. doi: 10.3390/s18082403.

[3] ZHANG Puning, KANG Xuyuan, WU Dapeng, *et al*. High-accuracy entity state prediction method based on deep belief network toward IoT search[J]. *IEEE Wireless Communications Letters*, 2019, 8(2): 492–495. doi: 10.1109/LWC.2018.2877639.

[4] ZHANG Puning, KANG Xuyuan, LIU Yuzhe, *et al*. Cooperative willingness aware collaborative caching mechanism towards cellular D2D communication[J]. *IEEE Access*, 2018, 6: 67046–67056. doi: 10.1109/ACCESS.2018.2873662.

[5] WU Dapeng, LIU Bingxu, YANG Qing, *et al*. Social-aware cooperative caching mechanism in mobile social networks[J]. *Journal of Network and Computer Applications*, 2020, 149: 102457. doi: 10.1016/j.jnca.2019.102457.

[6] 高云全, 李小勇, 方滨兴. 物联网搜索技术综述[J]. 通信学报, 2015, 36(12): 57–76. doi: 10.11959/j.issn.1000-436x.2015315.
GAO Yunquan, LI Xiaoyong, and FANG Binxing. Survey on the search of internet of things[J]. *Journal on Communications*, 2015, 36(12): 57–76. doi: 10.11959/j.issn.1000-436x.2015315.

[7] 张普宁, 刘元安, 吴帆, 等. 物联网中适用于内容搜索的实体状态匹配预测方法[J]. 电子与信息学报, 2015, 37(12): 2815–2820. doi: 10.11999/JEIT150191.
ZHANG Puning, LIU Yuan'an, WU Fan, *et al*. An entity state matching prediction method for content-based search in the internet of things[J]. *Journal of Electronics & Information Technology*, 2015, 37(12): 2815–2820. doi: 10.11999/JEIT150191.

[8] 邹宇驰, 刘松, 于楠, 等. 基于搜索的物联网设备识别框架[J]. 信息安全学报, 2018, 3(4): 25–40. doi: 10.19363/J.cnki.cn10-1380/tn.2018.07.03.
ZOU Yuchi, LIU Song, YU Nan, *et al*. IoT device recognition framework based on Web search[J]. *Journal of Cyber Security*, 2018, 3(4): 25–40. doi: 10.19363/J.cnki.cn10-1380/tn.2018.07.03.

[9] 李强, 贾煜璇, 宋金珂, 等. 网络空间物联网信息搜索[J]. 信息安全学报, 2018, 3(5): 38–53.
LI Qiang, JIA Yuxuan, SONG Jinke, *et al*. Search of internet of thing information in the cyberspace[J]. *Journal of Cyber Security*, 2018, 3(5): 38–53.

[10] MA Huadong and LIU Wu. A progressive search paradigm for the internet of things[J]. *IEEE MultiMedia*, 2018, 25(1): 76–86. doi: 10.1109/MMUL.2017.265091429.

[11] YAP K K, SRINIVASAN V, and MOTANI M. Max: Wide area human-centric search of the physical world[J]. *ACM Transactions on Sensor Networks*, 2008, 4(4): 26. doi: 10.1145/1387663.1387672.

[12] WANG Haodong, TAN C C, and LI Qun. Snoogle: A search engine for pervasive environments[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2010, 21(8): 1188–1202. doi: 10.1109/TPDS.2009.145.

[13] TAN C C, SHENG Bo, WANG Haodong, *et al*. Microsearch: A search engine for embedded devices used in pervasive computing[J]. *ACM Transactions on Embedded*

Computing Systems, 2010, 9(4): 43. doi: 10.1145/1721695.1721709.

[14] FATHY Y, BARNAGHI P, and TAFAZOLLI R. Large-scale indexing, discovery, and ranking for the Internet of Things (IoT)[J]. *ACM Computing Surveys*, 2018, 51(2): 29. doi: 10.1145/3154525.

[15] 刘强强, 余黎青, 赵鹏, 等. 基于移动平台的图像检索系统[J]. 计算机技术与发展, 2016, 26(11): 10–13.
LIU Qiangqiang, YU Liqing, ZHAO Peng, *et al*. A novel image retrieval system based on mobile platform[J]. *Computer Technology and Development*, 2016, 26(11): 10–13.

[16] TRUONG C, RÖMER K, and CHEN Kai. Fuzzy-based sensor search in the web of things[C]. The 3rd IEEE International Conference on the Internet of Things, Wuxi, China, 2012: 127–134. doi: 10.1109/IOT.2012.6402314.

[17] 刘素艳, 刘元安, 吴帆, 等. 物联网中基于相似性计算的传感器搜索[J]. 电子与信息学报, 2018, 40(12): 3020–3027.
LIU Suyan, LIU Yuanan, WU Fan, *et al*. Sensor search based on sensor similarity computing in the internet of things[J]. *Journal of Electronics & Information Technology*, 2018, 40(12): 3020–3027.

[18] LI Zhidu, JIANG Yuming, GAO Yuehong, *et al*. On buffer-constrained throughput of a wireless-powered communication system[J]. *IEEE Journal on Selected Areas in Communications*, 2019, 37(2): 283–297. doi: 10.1109/JSAC.2018.2872374.

[19] PATTAR S, BUYYA R, VENUGOPAL K R, *et al*. Searching for the IoT resources: Fundamentals, requirements, comprehensive review, and future directions[J]. *IEEE Communications Surveys & Tutorials*, 2018, 20(3): 2101–2132. doi: 10.1109/COMST.2018.2825231.

[20] ZHANG Zufan, ZENG Tian, YU Xiulan, *et al*. Social-aware D2D pairing for cooperative video transmission using matching theory[J]. *Mobile Networks and Applications*, 2018, 23(3): 639–649. doi: 10.1007/s11036-017-0973-z.

[21] Intel Berkeley Research Lab. Intel berkeley research lab sensors data[EB/OL]. http://db.csail.mit.edu/labdata/labdata.html, 2004.

张普宁：男，1988年生，博士，研究方向为物联网搜索.
亢旭源：男，1991年生，硕士生，研究方向为物联网搜索.
刘宇哲：男，1995年生，硕士生，研究方向为物联网搜索.
李学芳：女，1995年生，硕士生，研究方向为物联网搜索.
吴大鹏：男，1979年生，教授，研究方向为泛在无线网络、社会计算、互联网服务质量控制等.
王汝言：男，1969年生，教授，研究方向为泛在网络、全光网络理论与技术、多媒体信息处理等.

责任编辑：阮　望