基于主题模型的位置感知订阅发布系统

鲜学丰[1] 崔志明[1,2] 赵朋朋[2] 刘昭斌[1] 顾才东[1]
(江苏省现代企业信息化应用支撑软件工程技术研发中心 江苏 苏州 215104)[1]
(苏州大学智能信息处理及应用研究所 江苏 苏州 215006)[2]

摘 要 随着移动互联网的迅速发展和智能手机的普及,基于位置感知的订阅发布系统在工业界和学术界引起了广泛重视。现有系统主要处理海量空间数据下订阅与事件的查询匹配问题,其匹配模型主要是基于空间关键字之间的相似性,鲜有研究考虑语义相关性。为了探索并实现订阅发布系统在语义上的查询与匹配,提出了一种基于主题模型的位置感知订阅发布系统。首先,该系统利用主题模型对订阅发布系统中的关键字进行主题映射。然后,设计了一种两步分区索引结构 RP^{TM}-trees,并使用该索引结构为订阅的主题集合和空间信息建立索引。RP^{TM}-trees 根据主题集合的主题个数及关键主题对订阅进行两步分区索引,使其对订阅的分区能力更强,从而显著提升查询匹配的效率。最后,在高流速的事件流、千万级订阅数据集上进行了实验,实验结果表明所提方案是稳定和高效的。

关键词 订阅/发布,概率主题模型,主题映射,索引
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.03.026

Location-awareness Publication Subscription System Based on Topic Model

XIAN Xue-feng[1] CUI Zhi-ming[1,2] ZHAO Peng-peng[2] LIU Zhao-bin[1] GU Cai-dong[1]
(Jiangsu Province Support Software Engineering R&D Center for Modern Information Technology Application in Enterprise, Suzhou, Jiangsu 215104, China)[1]
(Institute of Intelligent Information Processing and Application, Soochow University, Suzhou, Jiangsu 215006, China)[2]

Abstract Location-awareness publication subscription system has drawn extensive academic and industrial attention with the booming development of mobile Internet and the popularity of smart-phones. The existing systems on location-awareness publication/subscription mainly focus on handling the query and matching problem of events among massive spatial data, whose matching model is mainly based upon the similarities of spatial keywords, while the semantic aspect is ignored. In order to explore how to realize the semantic query and matching in subscription/publication system, this paper proposed a location-awareness publication/subscription system based upon theme model. Firstly, the system makes use of theme model algorithm and realizes the thematic reflection of keywords in location-awareness publication/subscription system. Secondly, it designs a two-step partition index structure RP^{TM}-trees and utilizes RP^{TM}-trees to create an index between thematic aggregation and spatial information. As RP^{TM}-trees conducts a two-step partitioning and indexing of the subscription information based on the topic numbers of thematic aggregation and key topics, a stronger subscription partitioning ability is achieved, and the efficiency of query and matching is significantly improved. Finally, an experiment on high-speed event stream and millions and millions subscription data aggregation was conducted, indicating the effectiveness and the efficiency of the proposed solution.

Keywords Publication/Subscription, LDA, Topic mapping, Index

1 引言

随着移动互联网的迅速发展和智能手机(具有 GPS 功能)的普及,基于位置感知的订阅发布系统受到了越来越多工业界和学术界的关注。订阅者在订阅发布系统中提交感兴趣的订阅,发布者发布信息作为事件,如果订阅与事件匹配,那么事件会被推送给订阅者。基于位置感知的订阅发布系统已在许多主流的互联网产品上得到了广泛应用,如微博、Twitter

到稿日期:2016-12-28 返修日期:2017-04-17 本文受国家自然科学基金资助项目(61672372,61440053,61472268,61472211),江苏省高校"青蓝工程"优秀青年骨干教师培养项目,江苏高等学校优秀科技创新团队资助项目资助。

鲜学丰(1980-),男,博士,副教授,CCF 会员,主要研究方向为 Web 数据管理、数据挖掘,E-mail:sudaxxf@163.com;**崔志明**(1961-),男,硕士,教授,博士生导师,主要研究方向为智能信息处理新技术;**赵朋朋**(1980-),男,博士,副教授,硕士生导师,主要研究方向为空间数据处理,E-mail:ppzhao@suda.edu.cn(通信作者);**刘昭斌**(1965-),男,硕士,教授,主要研究方向为感知计算、无线传感器网络;**顾才东**(1963-),男,硕士,教授,硕士生导师,主要研究方向为智能信息处理和物联网。

以及一些同城交易平台。微博用户如果提交了一个包含文本兴趣描述和空间位置约束的订阅,如:“外滩附近新开咖啡厅”,那么“南京路 125 号咖啡厅开张,欢迎光临!”这条微博信息将会被作为事件推送给用户。此处的“外滩附近”被抽象为一个由经纬度范围表示的空间区域,“南京路 125 号”被抽象为一个空间点,如果该空间点落入订阅的空间区域且关键字匹配,那么该事件将会被推送给订阅者。

目前,学术界和工业界从不同侧面对基于位置感知的订阅发布系统进行了深入研究。主要有基于结构化数据[1-8]和非结构化数据[9-22]的位置感知订阅发布系统。在结构化数据方面:Sadoghi 等人[4]将一种布尔表达式索引应用到位置感知的订阅发布系统,同时将空间信息维度以谓词的形式加入到布尔表达式中,以实现位置感知的订阅发布系统;Jiang 等人[6]提出了一种 Ri-tree 索引树,该索引树能为每个事件返回 Top-k 个订阅;Guo 等人[7]提出了一种新的位置感知订阅发布系统,该系统能连续监控移动的订阅者,使其可以接收来自社交媒体和电子商务网站的结构化事件信息流。在非结构化数据方面:Li 等人[9]提出了一种高效的基于位置感知的订阅发布系统,该系统可在高流速的事件流、千万级订阅数据集上实现快速高效的检索;Yu 等人[20]在文献[9]的基础上进一步提出了为每个事件返回 Top-k 订阅的算法;Chen 等人[21]在位置感知的订阅发布系统中引入时间维度并设计了使每个订阅都能维持时新 Top-k 事件的查询匹配算法。

然而,在上述两个方面的工作中订阅与事件的匹配都是关键词匹配,未考虑语义级匹配。当用户订阅“咖啡”时,“星巴克”也应该能匹配,这符合用户订阅意图。因此,非常有必要将语义匹配引入基于位置感知的订阅发布系统中,以提升用户体验。本文将主题模型引入到位置感知订阅发布系统中,提出了一种新的基于主题模型的位置感知订阅发布系统(Location-Aware Publish Subscribe basing Topic Model,LP-STM),同时设计了一种高效的分区索引结构 RP^{TM}-trees。实验结果表明,该系统可以在高流速的事件流、千万级订阅数据集上实现快速且高效的匹配。

2 基于主题模型的位置感知订阅发布系统的问题描述

2.1 订阅与事件

1)订阅

在基于位置感知的订阅发布系统中,一个订阅 $s=\{s.T, s.R\}$ 由文本描述信息 $s.T$ 和空间区域信息 $s.R$ 两部分组成,文本描述信息是指订阅者描述需求的语言文字,$s.T$ 由一个关键字集合组成,即 $s.T=\{K_{s1},K_{s2},\cdots,K_{sn}\}$。空间信息是指订阅者所感兴趣的空间区域,对于空间区域 $s.R$,采用根据经纬度划分的最小边界矩形(MBR)来表示,例如:$(51.232256<lat<51.414526, 128.514434<long<127.534123)$。因此,订阅 s 可表示为:

$$s=\{[K_{s1},K_{s2},\cdots,K_{sn}],R\} \tag{1}$$

2)事件

一个事件 e 由一个文本信息集合 $e.T$ 和一个空间信息点 $e.loc$ 组成。$e.T$ 与 $s.T$ 的定义相似,即 $e.T=\{K_{e1},K_{e2},\cdots,K_{em}\}$,$e.loc$ 是一个由经纬度表示的空间信息点,例如:$(lat=51.332546, long=128.025464)$。因此,事件 e 可表示为:

$$e=\{[K_{e1},K_{e2},\cdots,K_{em}],loc\} \tag{2}$$

2.2 主题映射

本文在基于主题模型的位置感知订阅发布系统中采用了最常用的概率主题模型 LDA(Latent Dirichlet Allocation),使订阅与事件的关键字集合获得其主题映射。将订阅与事件的文本信息($s.T$ 和 $e.T$)作为 LDA 的训练集与验证集,通过不断地训练与验证,收敛得到关键字所对应的主题分布。对于一个订阅的关键字集合,从每个关键字 K_{si} 对应的主题 K_{si}^{TPC} 中选取分布概率最高的主题 K_{smax}^{TPC},然后将所有分布概率最高的主题的合取式作为订阅关键字集合的主题集合 $s.T_{TPC}$,即 $s.T_{TPC}=\{K_{smax_1}^{TPC}\wedge K_{smax_2}^{TPC}\wedge\cdots\wedge K_{smax_n}^{TPC}\}$。因此,订阅 s 可表示为:

$$s=\{[K_{smax_1}^{TPC}\wedge K_{smax_2}^{TPC}\wedge\cdots\wedge K_{smax_n}^{TPC}],R\} \tag{3}$$

事件的文本信息 $e.T$ 中关键字的主题映射获取方法与 $s.T$ 的处理方法相同,即 $e.T_{TPC}=\{K_{emax_1}^{TPC}\wedge K_{emax_2}^{TPC}\wedge\cdots\wedge K_{emax_m}^{TPC}\}$。因此,事件 e 可表示为:

$$e=\{[K_{emax_1}^{TPC}\wedge K_{emax_2}^{TPC}\wedge\cdots\wedge K_{emax_m}^{TPC}],loc\} \tag{4}$$

2.3 订阅与事件的匹配模式

基于主题模型的位置感知订阅发布系统的匹配模式有主题匹配、主题集合匹配及空间信息匹配 3 种。

定义 1(主题匹配) 对于一个给定的订阅主题 $K_{smax_i}^{TPC}$ 和事件主题 $K_{emax_j}^{TPC}$,根据匹配算法 M,如果 $M(K_{smax_i}^{TPC},K_{emax_j}^{TPC})>\&$($\&$ 为匹配阈值),那么主题 $K_{emax_j}^{TPC}$ 与主题 $K_{smax_i}^{TPC}$ 匹配。

定义 2(主题集合匹配) 对于一个给定的订阅主题集合 $s.T_{TPC}$ 和事件主题集合 $e.T_{TPC}$,根据匹配算法 M,如果对于 $\forall K_{smax_i}^{TPC}\in s.T_{TPC}$ 都有 $M(K_{smax_i}^{TPC},K_{emax_j}^{TPC})>\&$($\&$ 为匹配阈值,$K_{emax_j}^{TPC}\in e.T_{TPC}$),那么主题集合 $e.T_{TPC}$ 与主题集合 $s.T_{TPC}$ 匹配。

定义 3(空间信息匹配) 对于一个给定订阅的空间区域 $s.R$ 和事件的空间点 $e.loc$,如果空间点 $e.loc$ 落入空间区域 $s.R$,那么事件的空间信息点 $e.loc$ 与订阅的空间区域信息 $s.R$ 匹配。

以上给出了基于主题模型的位置感知的订阅发布系统的 3 种匹配的定义,接下来,本文给出基于主题模型的事件 e 与订阅 s 匹配的定义。

定义 4(订阅与事件匹配) 对于一个给定的订阅 s 与事件 e,如果事件 e 的主题集合 $e.T_{TPC}$ 与订阅 s 的主题集合 $s.T_{TPC}$ 匹配且事件 e 的空间信息点与订阅 s 的空间区域信息 $s.R$ 匹配,那么事件 e 与订阅 s 匹配。

最后,定义基于主题模型的位置感知订阅发布系统(LP-STM)需要解决的问题。

定义 5(LPSTM) 对于一个给定的事件流 E 和一个给定的订阅集合 S,基于主题模型的位置感知订阅发布系统的目的是找出与给定订阅 s 匹配的事件 e,其中 $s\in S, e\in E$。

为了便于理解,本文给出如下实例。给定 3 个订阅和 1 个事件,即 $s_1=\{$“KFC,可以美团。”,$R_1\}$,$s_2=\{$“麦当劳,上门

服务,酬宾活动。",R_2},s_3={"附近的必胜客,饿了么。",R_3},e_1={"东环路肯德基开张!欢迎食客光临,可送餐上门!",loc_1}。该实例的主题映射和空间信息分布如图 1 所示,其中 e_1 匹配 s_1,因为 s_1 的主题集合[快餐,外卖]被 e_1 的主题集合[快餐,外卖]匹配,且 loc_1 落入 R_1,两者的主题集合和空间信息都匹配。e_1 不匹配 s_2,虽然 loc_1 落入 R_2,但是 s_2 的主题集合[快餐,促销,外卖]与 e_1 的主题集合[快餐,外卖]不匹配,因此 e_1 不匹配 s_2。另外,e_1 不匹配 s_3,因为 loc_1 没有落入 R_2,两者的空间信息不匹配,因此 e_1 不匹配 s_3。

s/e TPC	快餐	促销	…	外卖	e.R/e.loc
e_1	肯德基	…	…	送餐	loc_1
s_1	KFC	…	…	美团	R_1
s_2	麦当劳	酬宾	…	上门	R_2
s_3	必胜客	…	…	饿了么	R_3
…	…	…	…	…	…

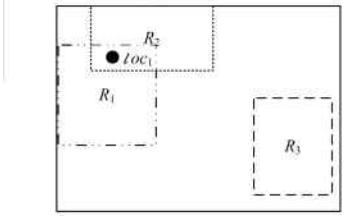

图 1　基于主题模型的位置感知订阅发布系统的订阅与事件示例

Fig. 1　Sample of subscriptions and events in LPSTM

3　基准解决方案

本节为基于主题模型的位置感知订阅发布系统的查询匹配问题提出了两个初步的解决方案,这两个解决方案将会被用作基准方案与优化的最终解决方案来进行实验对比。

基于主题模型的位置感知订阅发布系统的查询匹配有两个维度的问题需要处理:1)文本信息中关键字对应的主题集合匹配;2)空间信息匹配。本节提出的两个初步方案都是串行的,将主题集合维度和空间信息维度按顺序计算。方案的基本思路是在主题集合维度上,采用著名的倒排索引来索引订阅的主题集合,然后以空间信息索引树 R-tree 来索引订阅的空间信息,在该基本思路的基础上本文扩展出了两个初步的解决方案。方案 1:首先根据事件 e 的空间信息点从索引树 R-tree 上过滤出符合订阅 s 空间信息约束的候选订阅集合;然后通过订阅的主题集合倒排索引验证,获得与主题集合匹配的订阅。方案 2:首先通过订阅主题倒排索引获得发生主题集合匹配的候选集,然后在索引树R-tree上验证其空间信息维度是否匹配,如果空间信息匹配,则返回最终结果。本文称这两个方案分别为空间信息优先方法(S-Fist)和主题集合优先方法(TPC-First)。

4　基于 RP^{TM}-trees 的解决方案

为了提高订阅与事件的匹配效率,本文提出了一种新颖的分区索引结构 RP^{TM}-trees,在 RP^{TM}-trees 中每个订阅都有一个标志性主题,即关键主题 δ。δ 是指主题集合 $s.T_{TPC}$ 中在数据集上主题分布出现频率最小的主题。RP^{TM}-trees 首先根据订阅主题集合的主题个数 N 来分区所有订阅,然后根据关键主题 δ 对分区后的订阅子集进一步划分子集,通过上述两步将具有相同 N 值和相同关键主题 δ 的订阅划分到同一个子集中。为了索引空间信息,本文根据每个子集中订阅的空间区域信息 $s.R$ 建立若干 R-tree。给定一个事件 e,首先根据 $e.T_{TPC}$ 找到其匹配订阅子集的 R-tree,然后根据 $e.loc$ 获得空间信息匹配的订阅候选集,最后在主题集合索引上验证订阅候选集,获得匹配结果。

4.1　主题集合索引

RP^{TM}-trees 索引结构分两步索引订阅中的主题集合。第一步,根据它们的主题个数 N 将所有订阅划分为若干个互不相交的子集,表达式如下:

$$S=L_{(N1)}\cup L_{(N2)}\cup L_{(Ni)}\cup\cdots\cup L_{(Nn)} \tag{5}$$

如果 e 的主题集合匹配 s 的主题集合,那么 e 的主题个数一定大于或等于 s 的主题个数;如果不满足上述条件,则 s 中必然有一个主题没有出现在 e 的主题集合中,根据定义 2,e 肯定不是 s 的结果之一。第二步,对于具有相同个数的订阅,根据它们的关键主题,将其进一步进行分区,表达式如下:

$$L_{(Ni)}=L_{(\delta 1)}\cup L_{(\delta 2)}\cup L_{(\delta i)}\cup\cdots\cup L_{(\delta n)} \tag{6}$$

根据定义 2 可知,如果事件 e 匹配订阅 s,那么 s 中所有的主题都要包含在 e 的主题集合中。如果 s 的一个主题没有出现在 e 的主题集合中,那么 e 就不是 s 的结果。因此,对于事件 e,仅需考虑关键主题出现在 e 的主题集合中的订阅。在数据集中具有低频率的主题更具有过滤订阅的作用,因为较低频的主题出现在另一个事件中的可能性也较低。因此,本文选择订阅主题集合内出现频率最低的主题作为关键主题。

根据图 1 建立的主题集合索引如图 2 所示。第一步,根据订阅主题集合的主题个数 N 划分出两个子集 L_2 和 L_3。然后根据不同主题的出现频率选取出关键主题。在此,假设"快餐""促销"为数据集的关键主题。给定一个事件 e_1,其主题集合的主题个数为 2,根据定义 2,在 L_3 中的订阅与 e_1 不匹配。

s/e TPC	快餐	促销	…	外卖	e.R/e.loc
e_1	肯德基	…	…	送餐	loc_1
s_1	KFC	…	…	美团	R_1
s_2	麦当劳	酬宾	…	上门	R_2
s_3	必胜客	…	…	饿了么	R_3
…	…	…	…	…	…

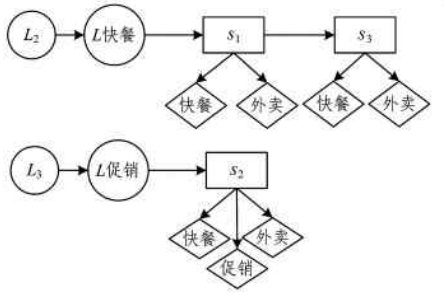

图 2　主题集合索引结构

Fig. 2　Topic collection index structure

4.2　基于 RP^{TM}-trees 索引结构的查询匹配

本节给出 RP^{TM}-trees 的索引结构。如图 3 所示,RP^{TM}-trees 由两部分组成:1)根据上文提出的两步分区法划分的两层主题集合倒排索引;2)根据相应订阅子集的空间区域信息

建立的空间索引树 R-tree。这里的 R-tree 是用来过滤事件的空间信息，并生成匹配订阅的候选集。基于 RP^{TM}-trees 的查询匹配过程为：给定一个事件 e，首先根据已学习到的主题模型找到关键字的主题映射，然后根据 $e.T_{TPC}$ 在 RP^{TM}-trees 索引上找到其匹配订阅子集的 R-tree，进一步根据 $e.loc$ 找到匹配空间信息的订阅候选集，最后在主题集合索引上验证这些订阅候选集，从而获得匹配结果。

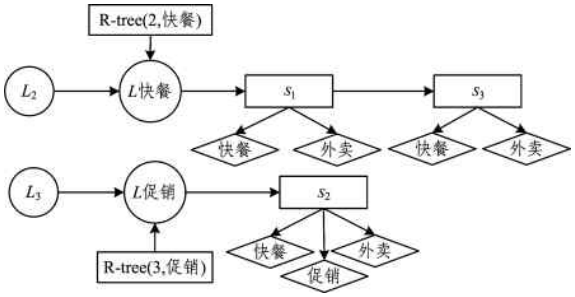

图 3 RPTM-trees 索引结构

Fig. 3 RPTM-trees index structure

本文使用图 1 的实例来解释基于 RP^{TM}-trees 的查询匹配过程。对于给定的事件 e_1={“东环路肯德基开张！欢迎食客光临，可送餐上门！”,loc_1}。首先根据已学习到的主题模型找到关键字的主题映射，即该事件可被重写为 e_1={[快餐 ∧ 外卖],loc_1}。由此可以计算出 e_1 的主题集合的主题个数为 2，由于 L_3 中的订阅有 3 个主题，因此首先被剪枝，接着继续搜索访问 L_2 中的订阅，发现 e_1 中存在主题“快餐”，然后根据 loc_1 检索相应的 R-tree(2，快餐)，发现 loc_1 落在 s_1 的空间区域 R_1 上，最后进一步验证 s_1 的主题集合中所有主题“快餐”和“外卖”都在 e_1 中的主题集合中出现，因此订阅 s_1 是事件 e_1 的最终匹配结果。

5 实验

为了验证本文提出的解决方案的有效性，首先分析 3 种索引方案的内存开销性能，然后在不同的订阅个数、不同的主题个数和不同的事件主题集合平均长度等参数变化下分析基于 RP^{TM}-trees 解决方案的性能，并与两组基准的方案进行对比。

实验中所有的索引结构都在内存中搭建，实验代码由 JAVA 编写。实验软硬件环境为：centos 5.6 服务器操作系统，256GB 内存，64kB L1 cache。

5.1 实验数据集和主题集合索引参数设置

本文将微博签到信息记录作为实验数据集。每个用户的签到信息包括：用户的 id、用户签到所揭示的空间位置(经纬度)以及用户的微博文本。这些微博可以作为事件信息流。另外，根据每个用户签到的空间信息点，将以一定长、宽随机生成的一个空间区域作为订阅者所订阅的空间区域，并将该微博数据作为订阅内容。由于微博数据量有限，本文根据标点将微博文本拆分为若干条文本信息，从而生成多种订阅。然后，根据微博的文本信息，采用主题模型 LDA 算法得到每个微博关键字的所有主题分布中分布概率最大的主题作为关键字对应的主题，从而得到每一个关键字的主题映射。实验总共生成了 1000 万个订阅和 10 万个事件作为匹配测试数据。表 1 详细介绍了数据集的参数设置。

表 1 参数设置

Table 1 Parameter settings

参数	微博
订阅的个数	2M,4M,6M,8M,10M
订阅主题集合的平均长度	4～8
事件主题集合的平均长度	5～10
主题的齐夫分布	0.2,0.4,0.6,0.8,1.0
整个数据集中主题的个数	200,400,600,800,1000
R-tree 的节点容量	40

5.2 结果分析

实验将比较 RP^{TM}-trees 和两种基准方案 S-Fist，TPC-First 的性能。由于这 3 种索引都是内存索引，因此实验首先分析了这 3 种索引的内存开销情况，然后分别以不同的订阅个数、不同的主题个数以及不同的事件主题集合的平均长度参数进行实验对比，以分析它们的性能。

5.2.1 内存开销

首先对比 3 种索引方案随着订阅数量的增长引起的内存开销变化，实验结果如图 4 所示。从图 4 中可以看出，3 种索引的内存开销随着订阅数量的增长而增长，在固定订阅数量不变的情况下，三者的内存开销几乎一致，相较于 S-Fist 和 TPC-First，RP^{TM}-trees 的内存开销较大。经分析这是因为相较于 S-Fist 和 TPC-First 索引，RP^{TM}-trees 产生了更多的 R-tree，这将导致内存开销增大，但这并未显著增加内存开销，仅为略微增大。S-Fist 和 TPC-First 的索引结构是一样的，区分这两个解决方案的主要方法是判断在查询方案上采用的是空间信息优先(S-Fist)还是主题集合信息优先(TPC-First)。

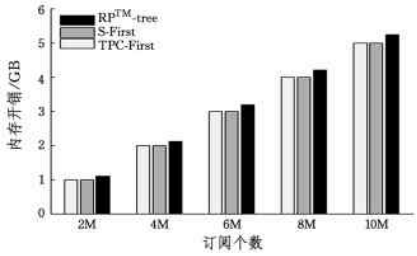

图 4 内存开销

Fig. 4 Memory overhead

5.2.2 不同的订阅个数

为了测试这 3 种方案的稳定性，我们在不同订阅数量的分布下进行实验，平均事件匹配时间在不同订阅数量下的分布如图 5 所示。

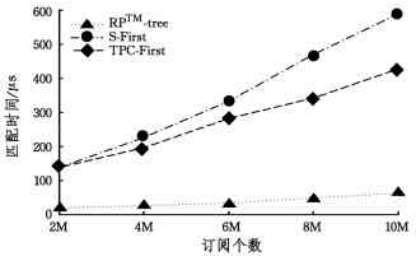

图 5 订阅个数

Fig. 5 Number of subscriptions

从图 5 可知，在绝对的平均匹配时间和索引的稳定性两方面，RP^{TM}-trees 的表现都是最好的，其次是 TPC-First。其原因在于 RP^{TM}-trees 根据主题集合长度大小以及关键主题对订阅的分区能力远高于后两者。另外，由于 RP^{TM}-trees 对订阅的分区能力更强，使得每个相应的R-tree索引的空间信息量大幅减少，从而提高了 R-tree 对订阅的过滤效率。

5.2.3　不同的主题个数

对于 3 种索引，主题个数是一个非常重要的参数，因为 3 种索引都是全部或部分根据主题个数划分订阅子集的。如图 6 所示，当主题数量增加时，3 种索引的平均事件匹配时间都在减少，这是因为随着主题数量的增加，3 种索引都会产生更小的订阅分区，而 RP^{TM}-trees 减小得更明显，这是因为 RP^{TM}-trees 首先根据关键主题划分订阅，当主题个数增加时，单个索引的大小将会明显减少，结合 R-tree 在空间信息上的过滤功能，使得事件匹配时间相较于 S-Fist，TPC-First 进一步减少。另外，随着数据集中主题数量的增加，事件匹配订阅的可能性将进一步增加，这是因为随着主题数量的增加，订阅和事件的相关性将增大。

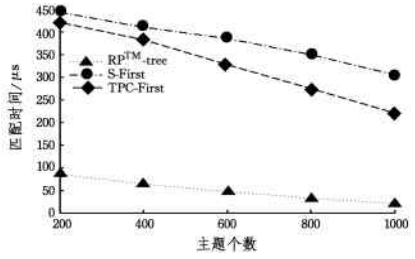

图 6　主题个数

Fig. 6　Number of topics

5.2.4　不同事件主题集合长度

该实验结果如图 7 所示，在相同的主题集合长度的情况下，RP^{TM}-trees 的匹配时间明显优于 S-Fist 和 TPC-First。但从图 7 中也可以得出，3 种索引中只有 RP^{TM}-trees 对事件的主题长度参数较为敏感。因为相较于 S-Fist 和 TPC-First，RP^{TM}-trees 采用主题集合的长度分区订阅，随着事件的主题集合长度的不断增大，RP^{TM}-trees 根据主题集合长度剪枝的能力将减弱，使得平均事件匹配时间随着事件主题集合长度的增加而增加。但即便如此，在实验中主题集合长度最大为 10 时，RP^{TM}-trees 的匹配时间也显著优于 S-Fist 和 TPC-First。

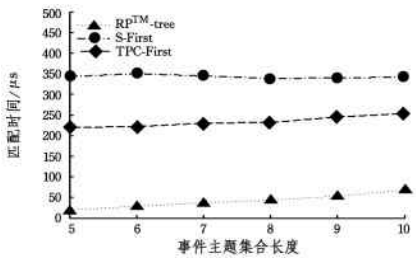

图 7　事件主题集合长度

Fig. 7　Length of event theme collection

结束语　本文分析了在位置感知的订阅发布系统上引入语义匹配的必要性，并提出了将主题模型应用到位置感知的订阅发布系统上的方案。同时，为基于主题模型的位置感知订阅发布系统设计了高效的索引结构以及基于该索引结构的查询匹配方法。实验结果表明，系统可以在高流速的事件流、千万级订阅数据集上实现快速且高效的匹配。如何将语义匹配引入到位置感知的订阅发布系统中将成为未来的重要研究方向，我们下一步将研究在基于主题模型的位置感知订阅发布系统上引入 Top-k 算法，使每个订阅都能维持时、空、语义等多个维度的 k 个时新的事件集合，并研究订阅结果是否满足用户的订阅意图和订阅系统用户满意度评估问题。

参 考 文 献

[1] CUGOLA G, MARGARA A. High-Performance Location-Aware Publish-Subscribe on GPUs[C]// International MIDDLEWARE Conference. Springer-Verlag New York, 2012: 312-331.

[2] OOI B C, TAN K L, TUNG A. Sense the physical, walkthrough the virtual, manage the co (existing) spaces: a database perspective[J]. Acm Sigmod Record, 2010, 38(3): 5-10.

[3] HU J, CHENG R, WU D, et al. Efficient Top-k Subscription Matching for Location-Aware Publish/Subscribe[C]// International Symposium on Spatial and Temporal Databases. Springer International Publishing, 2015: 333-351.

[4] SADOGHI M, JACOBSEN H A. Location-based matching in publish/subscribe revisited[C]// Proceedings of the Posters and Demo Track. ACM, 2012: 1-2.

[5] GUO L, ZHANG D, LI G, et al. Location-Aware Pub/Sub System: When Continuous Moving Queries Meet Dynamic Event Streams[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2015: 843-857.

[6] JIANG H, ZHAO P, SHENG V S, et al. An Efficient Location-Aware Publish/Subscribe Index with Boolean Expressions[C]// Web Information Systems Engineering-WISE 2015. Springer International Publishing, 2015.

[7] GUO L, CHEN L, ZHANG D, et al. Elaps: An efficient locationaware pub/sub system[C]// IEEE 31st International Conference on Data Engineering (ICDE). IEEE, 2015: 1504-1507.

[8] JIANG H H, ZHAO P P, SHENG V S, et al. AnEfficient Location-Aware Top-k Subscription Matching for Publish/Subscribe with Boolean Expressions [C]// International Conference on Database Systems for Advanced Applications. Springer International Publishing, 2016: 335-350.

[9] LI G, WANG Y, WANG T, et al. Location-aware publish/subscribe[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 802-810.

[10] LAI S A, WANG G Z. P2P Streaming Media Resource Location

Algorithm Based on publish/subscribe[J]. Henan Science, 2012,32(2):239-243. (in Chinese)
来杜安,王桂芝. 基于发布订阅的 P2P 流媒体资源定位算法[J]. 河南科学,2012,30(2):239-243.

[11] HUANG Y,GARCIAMOLINA H. Publish/Subscribe in a Mobile Environment[J]. Wireless Networks,2001,10(6):27-34.

[12] NAICKEN,MURUGAPA S. Trusted content-based publish/subscribe trees[D]. University of Sussex,2012.

[13] EUGSTER P T,GARBINATO B,HOLZER A. Location-based Publish/Subscribe[C]// IEEE International Symposium on Network Computing & Applications. 2005:279-282.

[14] HU H,LIU Y,LI G,et al. A location-aware publish/subscribe framework for parameterized spatio-textual subscriptions[C]// International Conference on Data Engineering. IEEE,2015:711-722.

[15] ZHENG K,SU H,ZHENG B,et al. Interactive Top-k Spatial Keyword queries[C]// IEEE 31st International Conference on Data Engineering (ICDE). IEEE,2015:423-434.

[16] ZHENG K,FUNG P C,ZHOU X. K-Nearest Nei-ghbor Search for Fuzzy Objects[C]// Association for Computing Machinery. Special Interest Group on Management of Data. International Conference Proceedings. Association for Computing Machinery (ACM),2010:699-710.

[17] LIU H,UNIVERSITY X,XIANGTAN. LCESM: Locati-on-aware Context Event Subscription Mechanism[J]. Computer Science,2011,38(7):80-79.

[18] WANG X,ZHANG Y,ZHANG W,et al. AP-Tree: Efficiently support continuous spatial-keyword queries over stream[C]// International Conference on Data Engineering. IEEE, 2015:1107-1118.

[19] PODNAR I. Location-aware Content Delivery Service Using Publish/Subscribe[OL]. http://people. kth. se/~devlic/publications/tcmc03_final. pdf.

[20] YU M,LI G,WANG T,et al. Efficient Filtering- Algorithms for Location-Aware Publish/Subscribe[J]. IEEE Transactions on Knowledge & Data Engineering,2015,27(4):950-963.

[21] CHEN L,CONG G,CAO X,et al. Temporal Spatial-Keyword-Top-k publish/subscribe[C]// International Conference on Data Engineering. IEEE,2015:255-266.

[22] MISHRA S,MURTHY C S R. An efficient location aware distributed physical resource block assignment for denseclosed access femtocell networks[J]. Computer Networksthe International Journal of Computer & Telecommunications Networking, 2016,94(C):164-175.

(上接第 130 页)

[14] LIN H L,LI Y,WANG W P,et al. Efficient segment pattern based method for malicious URL detection[J]. Journal on Communications,2015,36(Z1):141-148. (in Chinese)
林海伦,李焱,王伟平,等. 高效的基于段模式的恶意 URL 检测方法[J]. 通信学报,2015,36(Z1):141-148.

[15] YANG Z M,LI Q,LIU J R,et al. Research of Threat Intelligence Sharing and Using for Cyber Attack Attribution[J]. Journal of Information Securyity Research,2015,1(1):31-36. (in Chinese)
杨泽明,李强,刘俊荣,等. 面向攻击溯源的威胁情报共享利用研究[J]. 信息安全研究,2015,1(1):31-36.

[16] SAMTANI S,CHINN K,LARSON C,et al. AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis [C]// 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE,2016:19-24.

[17] AHREND J M,JIROTKA M,JONES K. On the collaborative practices of cyber threat intelligence analysts to develop and utilize tacit Threat and Defence Knowledge[C]// 2016 International Conference on Cyber Situational Awareness,Data Analytics And Assessment (CyberSA). IEEE,2016:1-10.

[18] DAI W,JI W. A mapreduce implementation of C4. 5 decision tree algorithm[J]. International Journal of Database Theory and Application,2014,7(1):49-60.

[19] PATIL T R,SHEREKAR S S. Performance analysis of Naive Bayes and J48 classification algorithm for data classification[J]. International Journal of Computer Science and Applications, 2013,6(2):256-261.

[20] PAN W,CHEN G. A method of off-line signature verification for digital forensics[C]// 2016 12th International Conference on Natural Computation,Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE,2016:488-493.

[21] VLADIMIR V N,VAPNIK V. The nature of statistical learning theory[M]. New York:Springer-verlag,1995:988-999.

[22] CRAMMER K,DREDZE M,PEREIRA F. Exact convex confidence-weighted learning[C]// Advances in Neural Information Processing Systems. 2009:345-352.

[23] HOI S C H,WANG J,ZHAO P. LIBOL: A Library for Online Learning Algorithms[J]. Journal of Machine Learning Research,2014,15(1):495-499.